

日 本 国 特 許 庁  
JAPAN PATENT OFFICE

11002 U.S. PTO  
10/084540  
02/28/02

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office

出 願 年 月 日

Date of Application:

2001年11月12日

出 願 番 号

Application Number:

特願2001-345523

[ ST.10/C ]:

[ JP2001-345523 ]

出 願 人

Applicant(s):

株式会社日立製作所

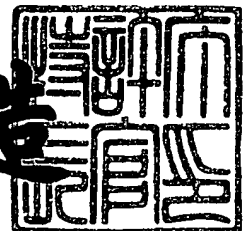
U.S. Appln Filed 2-28-02  
Inventor: K magi et al  
mattingly stangers malor  
Docket ASA-1072

CERTIFIED COPY OF  
PRIORITY DOCUMENT

2002年 1月11日

特 許 庁 長 官  
Commissioner,  
Japan Patent Office

及 川 耕 造



【書類名】 特許願

【整理番号】 K01011371A

【あて先】 特許庁長官

【国際特許分類】 G06F 17/30

【発明者】

    【住所又は居所】 神奈川県川崎市麻生区王禅寺 1 0 9 9 番地 株式会社日立製作所 システム開発研究所内

    【氏名】 茂木 和彦

【発明者】

    【住所又は居所】 神奈川県川崎市麻生区王禅寺 1 0 9 9 番地 株式会社日立製作所 システム開発研究所内

    【氏名】 大枝 高

【発明者】

    【住所又は居所】 千葉県松戸市二十世紀が丘丸山町 1 7

    【氏名】 喜連川 優

【特許出願人】

    【識別番号】 000005108

    【氏名又は名称】 株式会社日立製作所

【代理人】

    【識別番号】 100075096

    【弁理士】

    【氏名又は名称】 作田 康夫

【手数料の表示】

    【予納台帳番号】 013088

    【納付金額】 21,000円

【提出物件の目録】

    【物件名】 明細書 1

    【物件名】 図面 1

    【物件名】 要約書 1

【プルーフの要否】 要

【書類名】 明細書

【発明の名称】 データベース管理システム情報を考慮したデータ再配置方法およびデータ再配置を行う計算機システム

【特許請求の範囲】

【請求項 1】

データベース管理システムが稼動する少なくとも 1 台以上の計算機と、前記データベース管理システムにより管理されるデータベースデータを格納した少なくとも 1 台以上の記憶装置と、前記計算機と前記記憶装置との間に接続され、前記計算機と前記記憶装置との間でデータの転送を制御する記憶制御手段と、前記記憶装置における前記データの配置を管理するデータ位置管理サーバとを有する計算機システムにおけるデータの再配置方法において、

前記データ位置管理サーバが、前記計算機から前記データベース管理システムにより管理されるデータベースに関する情報を取得し、

前記データ位置管理サーバが、前記データベースに関する情報を含む取得情報に基づいて、前記データベースデータの前記計算機システム内での配置を決定し

前記データ位置管理サーバから前記決定されたデータ配置を実現するためのデータ移動を前記記憶制御手段に指示し、

前記記憶制御手段は、前記指示に従って、前記記憶装置内に格納された前記データの配置を変更することを特徴とするデータの再配置方法。

【請求項 2】

前記データ位置管理サーバは、前記記憶制御手段からデータの記憶位置に関する情報を取得することを特徴とする請求項 1 記載のデータの再配置方法。

【請求項 3】

前記データ位置管理サーバの機能が、所定の前記データベースが稼動する計算機上で実現されることを特徴とする請求項 1 記載のデータの再配置方法。

【請求項 4】

前記データ位置管理サーバの機能が、所定の前記記憶制御手段上で実現されることを特徴とする請求項 1 記載のデータの再配置方法。

【請求項 5】

前記データ位置管理サーバの機能が、所定の前記記憶装置上で実現されることを特徴とする請求項1記載のデータの再配置方法。

【請求項 6】

前記記憶制御手段が、前記計算機上で実施されるプログラムにより実現されることを特徴とする請求項 1 記載のデータの再配置方法。

【請求項 7】

前記データ位置管理サーバから前記記憶制御手段に決定されたデータの配置を指示する前に、前記決定されたデータの配置を管理者に提示し、管理者にデータの配置変更を実施するか確認することを特徴とする請求項 1 記載のデータの再配置方法。

【請求項 8】

前記データベースに関する情報は、前記データベース管理システムのスキーマにより定義される表・索引・ログを含むデータ構造に関する情報と、前記データベースデータを前記スキーマにより定義されるデータ構造毎に分類した前記記憶装置における記録位置に関する情報の少なくとも1つを含むことを特徴とする請求項1に記載のデータの再配置方法。

【請求項 9】

前記記憶装置が複数台存在する前記計算機システムにおいて、前記データ位置管理サーバは、前記データの配置を決定する際に、前記データを既に記憶されている前記記憶装置とは異なる前記記憶装置に割り当てる前記データの配置を決定することを特徴とする請求項 8 に記載のデータの再配置方法。

【請求項 1 0】

前記記憶装置は少なくとも 1 つ以上のデータを記憶する物理記憶手段を有し、  
前記データ位置管理サーバは、前記データの配置を決定する際に前記物理記憶手段における記憶位置まで定めたデータの配置を決定することを特徴とする請求項 8 に記載のデータの再配置方法。

【請求項 1 1】

前記記憶装置のうち少なくとも 1 台以上は、少なくとも 1 つ以上のデータを記

憶する物理記憶手段を有し、

前記計算機が前記記憶装置をアクセスする際に利用する論理位置を前記物理記憶手段の記憶位置へ変換する論理－物理位置変換手段を有し、

前記データ位置管理サーバは、前記論理－物理位置変換手段を有する記憶装置から論理位置－物理位置のマッピングに関する記憶装置内論理－物理マッピング情報を含む情報を取得することを特徴とする請求項 1 0 に記載のデータの再配置方法。

【請求項 1 2】

前記論理－物理位置変換手段を有する記憶装置の少なくとも 1 台以上は、前記論理位置に対応するデータの物理記憶手段の記憶位置を変更する記憶装置内データ物理記憶位置変更手段を有し、

前記データ位置管理サーバは、前記決定のデータ配置を実現するためのデータ移動を指示する際に、前記記憶装置内データ物理記憶位置変更手段を有する記憶装置内のデータ移動に関しては前記記憶装置に対してデータ移動を指示し、

前記記憶装置は、前記指示に従って、前記記憶装置内に格納された前記データの配置を変更することを特徴とする請求項 1 1 に記載のデータの再配置方法。

【請求項 1 3】

前記データ位置管理サーバは、前記取得情報を基づいて、同時にアクセスされる可能性が高い前記データベースデータの組を検出し、前記組を異なる前記物理記憶手段に配置することを特徴とする請求項 1 0 に記載のデータの再配置方法。

【請求項 1 4】

前記データ位置管理サーバが、前記データベース管理システムにおける処理の実行履歴に関する情報を含む情報を取得することを特徴とする請求項 1 3 に記載の再配置方法。

【請求項 1 5】

前記記憶装置の少なくとも 1 台は前記記憶装置内の前記物理記憶手段の稼動情報を取得し、

前記データ位置管理サーバは、前記記憶装置から前記物理記憶稼動情報を取得し、

前記取得情報は、前記記憶装置から取得する前記物理記憶稼動情報を含むことを特徴とする請求項 1 3 に記載のデータの再配置方法。

【請求項 1 6】

前記データ位置管理サーバが、表データと前記表データに対する索引データを把握し、前記索引データが木構造のデータ構造の場合に前記索引データを前記同時にアクセスされる可能性が高いデータベースデータの組とすることを特徴とする請求項 1 3 に記載のデータの再配置方法。

【請求項 1 7】

前記データ位置管理サーバが、前記データベース管理システムにおけるデータの更新処理時に記録するログデータとその他の前記データベースデータを前記同時にアクセスされる可能性が高いデータベースデータの組とすることを特徴とする請求項 1 3 に記載のデータの再配置方法。

【請求項 1 8】

前記データベースに関する情報に、前記データベース管理システムが前記スキーマにより定義される同一のデータ構造に属する前記データベースデータをアクセスする際の並列度に関する情報を含み、

前記データ位置管理サーバは、前記取得情報に基づいて、前記スキーマにより定義される同一のデータ構造に属する前記データベースデータを複数の前記物理記憶手段に配置することを特徴とする請求項 1 0 に記載のデータの再配置方法。

【請求項 1 9】

前記データ位置管理サーバは、前記取得情報に基づいて、前記データベース管理システムが前記データベースデータをシーケンシャルにアクセスする際のアクセス場所とアクセス順を判断し、前記シーケンシャルにアクセスされるデータベースデータを前記物理記憶手段上で前記連続アクセス順の前後関係を保持しながら連続した領域に配置することを特徴とする請求項 1 0 に記載のデータの再配置方法。

【請求項 2 0】

前記データベースに関する情報は、前記データベース管理システムが前記データベースデータを前記計算機上のメモリにキャッシュする際のキャッシュメモリ

量とキャッシュ動作に関する情報を含み、

前記データ位置管理サーバは、前記記憶装置からキャッシュメモリに関する記憶装置キャッシュメモリ情報を取得し、

前記データベースに関する情報と前記記憶装置キャッシュメモリ情報から得られるキャッシュメモリの効果に関する情報に基づいて前記記憶装置にデータを配置することを特徴とする請求項 8 に記載のデータの再配置方法。

【請求項 2 1】

データベース管理システムが稼動する少なくとも 1 台以上の計算機と、前記データベース管理システムにより管理されるデータベースデータを格納した少なくとも 1 台以上の記憶装置と、前記計算機と前記記憶装置との間に接続され、前記計算機と前記記憶装置との間でデータの転送を制御する記憶制御手段と、前記記憶装置における前記データの配置を管理するデータ位置管理サーバとを有する計算機システムにおいて、

前記データ位置管理サーバは、前記計算機から前記データベース管理システムにより管理されるデータベースに関する情報を取得する情報取得手段と、

前記データベースに関する情報を含む取得情報に基づいて、前記データベースデータの前記計算機システム内での配置を決定する配置決定手段と、

前記データ位置管理サーバから前記記憶制御手段により決定されたデータの配置を実現するためのデータ移動を指示するデータ配置指示手段を有し、

前記記憶制御手段は、前記データ位置管理サーバからの指示に従って、前記複数の記憶装置内に格納された前記データの配置を変更するデータ配置変更手段を有することを特徴とする計算機システム。

【請求項 2 2】

前記データ位置管理サーバは、前記記憶制御手段からデータの記憶位置に関する情報を取得するデータ位置取得手段を有することを特徴とする請求項 2 1 記載の計算機システム。

【請求項 2 3】

前記データ位置管理サーバと所定の前記データベースが稼動する計算機が同一の計算機であることを特徴とする請求項 2 1 記載の計算機システム。



【請求項 2 4】

前記データ位置管理サーバと前記記憶制御手段が同一の装置であることを特徴とする請求項 2 1 記載の計算機システム。

【請求項 2 5】

前記データ位置管理サーバと所定の前記記憶装置が同一の装置であることを特徴とする請求項 2 1 記載の計算機システム。

【請求項 2 6】

前記記憶制御手段が、前記計算機上で実施されるプログラムにより実現されることを特徴とする請求項 2 1 記載の計算機システム。

【請求項 2 7】

前記配置決定手段により決定したデータの配置を管理者に提示する手段と、前記管理者からデータの配置変更の可否を取得する手段を有することを特徴とする請求項 2 1 記載の計算機システム。

【請求項 2 8】

前記データベースに関する情報は、前記データベース管理システムのスキーマにより定義される表・索引・ログを含むデータ構造に関する情報と、前記データベースデータを前記スキーマにより定義されるデータ構造毎に分類した前記記憶装置における記録位置に関する情報の少なくとも1つを含むことを特徴とする請求項 2 1 に記載の計算機システム。

【請求項 2 9】

前記記憶装置が複数台存在し、前記配置決定手段が、既に前記データが記憶されている前記記憶装置とは異なる前記記憶装置に割り当てる前記データの配置を決定することを特徴とする請求項 2 8 に記載の計算機システム。

【請求項 3 0】

前記記憶装置は少なくとも1つ以上のデータを記憶する物理記憶手段を有し、

前記配置決定手段は、前記記憶装置内の前記物理記憶手段における記憶位置に基づいてデータの配置を決定することを特徴とする請求項 2 8 に記載の計算機システム。

【請求項 3 1】

前記記憶装置のうち少なくとも1台以上は、前記計算機が前記記憶装置をアクセスする際に利用する論理位置を前記物理記憶手段の記憶位置へ変換する論理－物理位置変換手段を有し、

前記データ位置管理サーバは、前記論理－物理位置変換手段を有する記憶装置から論理位置－物理位置のマッピングに関する記憶装置内論理－物理マッピング情報を含む情報を取得することを特徴とする請求項30に記載の計算機システム

【請求項32】

前記論理－物理位置変換手段を有する記憶装置の少なくとも1台以上は、前記論理位置に対応するデータの前記物理記憶手段の記憶位置を変更する記憶装置内データ物理記憶位置変更手段を有し、

前記データ位置管理サーバが、前記記憶装置内データ物理記憶位置変更手段を有する記憶装置対して前記データの配置を実現するためのデータ移動を指示する記憶装置データ配置指示手段を有することを特徴とする請求項31に記載の計算機システム。

【請求項33】

前記配置決定手段が、前記取得情報に基づいて、同時にアクセスされる可能性が高い前記データベースデータの組を検出し、前記組を異なる前記物理記憶手段に配置することを特徴とする請求項30に記載の計算機システム。

【請求項34】

前記データベースに関する情報は、前記データベース管理システムにおける処理の実行履歴に関する情報を含むことを特徴とする請求項33に記載の計算機システム。

【請求項35】

前記記憶装置の少なくとも1台は前記記憶装置内の前記物理記憶手段の稼動情報を取得する稼動情報取得手段を有し、

前記データ位置管理サーバは、前記記憶装置から前記物理記憶手段稼動情報を取得する手段を有し、

前記配置決定手段が利用する前記取得情報に前記記憶装置から取得した前記物

理記憶手段稼動情報を含むことを特徴とする請求項 3 3 に記載の計算機システム

【請求項 3 6】

前記配置決定手段が、表データと前記表データに対する索引データを把握し、前記索引データが木構造のデータ構造の場合に前記索引データを前記同時にアクセスされる可能性が高いデータベースデータの組とすることを特徴とする請求項 3 3 に記載の計算機システム。

【請求項 3 7】

前記配置決定手段が、前記データベース管理システムにおけるデータの更新処理時に記録するログデータとその他の前記データベースデータを前記同時にアクセスされる可能性が高いデータベースデータの組とすることを特徴とする請求項 3 3 に記載の計算機システム。

【請求項 3 8】

前記データベースに関する情報は、前記データベース管理システムが前記スキーマにより定義される同一のデータ構造に属する前記データベースデータをアクセスする際の並列度に関する情報を含み、

前記配置決定手段が、前記取得情報に基づいて、前記スキーマにより定義される同一のデータ構造に属する前記データベースデータを複数の前記物理記憶手段に配置することを特徴とする請求項 3 0 に記載の計算機システム。

【請求項 3 9】

前記配置決定手段が、前記取得情報に基づいて、前記データベース管理システムが前記データベースデータをシーケンシャルにアクセスする際のアクセス場所とアクセス順を判断し、前記シーケンシャルにアクセスされるデータベースデータを前記物理記憶手段上で前記連続アクセス順の前後関係を保持しながら連続した領域に配置することを特徴とする請求項 3 0 に記載の計算機システム。

【請求項 4 0】

前記データベースに関する情報は、前記データベース管理システムが前記オブジェクトのデータを前記計算機上のメモリにキャッシュする際のキャッシュメモリ量とキャッシュ動作に関する情報を含み、

前記データ位置管理サーバは、前記記憶装置からキャッシュメモリに関する記憶装置キャッシュメモリ情報を取得する手段を有し、

前記配置決定手段が、前記データベースに関する情報と前記記憶装置キャッシュメモリ情報から得られるキャッシュメモリの効果に関する情報に基づいて前記記憶装置にデータを配置することを特徴とする請求項 2 8 に記載の計算機システム。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】

本発明は、データベース管理システムに関する。

【0002】

【従来の技術】

現在、データベース（DB）を基盤とする多くのアプリケーションが存在し、DBに関する一連の処理・管理を行うソフトウェアであるデータベース管理システム（DBMS）は極めて重要なものとなっている。特に、DBMSの処理性能はDBを利用するアプリケーションの性能も決定するため、DBMSの処理性能の向上は極めて重要である。

【0003】

DBの特徴の1つは、多大なデータを扱うことである。そのため、DBMSの実行環境の多くにおいては、DBMSが実行される計算機に対して大容量の記憶装置あるいは複数の記憶装置を接続し、記憶装置上にDBのデータを記憶する。そのため、DBに関する処理を行う場合に、記憶装置に対してアクセスが発生し、記憶装置におけるデータアクセス性能がDBMSの性能を大きく左右する。そのため、DBMSが稼動するシステムにおいて、記憶装置の性能最適化が極めて重要であり、特にDBMSが管理するデータの物理記憶装置に対する配置の最適化が重要である。

文献“Oracle 8 i パフォーマンスのための設計およびチューニング、リリース 8. 1”（日本オラクル社、部品番号 J 0 0 9 2 1 - 0 1）の第 2 0 章（文献 1）は、RDBMSである Oracle 8 i における I/O のチューニング

を述べている。その中で、RDBMSの内部動作のチューニングと共に、データの配置のチューニングに関連するものとして、ログファイルは他のデータファイルから分離した物理記憶装置に記憶すること、ストライプ化されたディスクにデータを記憶することによる負荷分散が効果があること、表のデータとそれに対応する索引データは異なる物理記憶装置に記憶すると効果があること、RDBMSとは関係ないデータを異なる物理記憶装置に記憶することが述べられている。

## 【 0 0 0 4 】

米国特許 6 0 3 5 3 0 6 (文献 2) は、DBMS - ファイルシステム - ボリュームマネージャー記憶装置間のマッピングを考慮した性能解析ツールに関する技術を開示している。この性能解析ツールは、各レイヤにおけるオブジェクトの稼動状況を画面に表示する。このときに上記のマッピングを考慮し、その各オブジェクトに対応する他レイヤのオブジェクトの稼動状況を示す画面を容易に表示する機能を提供する。また、ボリュームマネージャレイヤのオブジェクトに関して、負荷が高い記憶装置群に記憶されているオブジェクトのうち、2 番目に負荷が高いオブジェクトを、もっとも負荷が低い記憶装置群に移動するオブジェクト再配置案を作成する機能を有している。

## 【 0 0 0 5 】

特開平 9 - 2 7 4 5 4 4 号公報 (文献 3) は、計算機がアクセスするために利用する論理的記憶装置を、実際にデータを記憶する物理記憶装置に配置する記憶装置において、前記論理的記憶装置の物理記憶装置への配置を動的に変更することにより記憶装置のアクセス性能を向上する技術を開示している。アクセス頻度が高い物理記憶装置に記憶されているデータの一部を前記の配置動的変更機能を用いて他の物理記憶装置に移動することにより、特定の物理記憶装置のアクセス頻度が高くなるようにし、これにより記憶装置を全体としてみたときの性能を向上させる。また、配置動的変更機能による高性能化処理の自動実行方法についても開示している。

## 【 0 0 0 6 】

特開 2 0 0 1 - 6 7 1 8 7 号公報 (文献 4) は、計算機がアクセスするために利用する論理的記憶装置を、実際にデータを記憶する物理記憶装置に配置し、前

記論理的記憶装置の物理記憶装置への配置を動的に変更する機能を有する記憶装置において、論理的記憶装置の物理記憶装置への配置の変更案を作成する際に、物理記憶装置を属性の異なるグループに分割し、それを考慮した配置変更案を作成し、その配置変更案に従って自動的に論理的記憶装置の配置を変更する技術を開示している。配置変更案作成時に、物理記憶装置を属性毎にグループ化し、論理的記憶装置の配置先として、それが有している特徴にあった属性を保持している物理記憶装置のグループに属する物理記憶装置を選択する配置変更案を作成することにより、良好な配置変更案を作成する。

【 0 0 0 7 】

【発明が解決しようとする課題】

従来の技術には以下のような問題が存在する。

【 0 0 0 8 】

文献 1 で述べられているものは、管理者がデータの配置を決定する際に考慮すべき項目である。現在、1 つの記憶装置内に多数の物理記憶装置を有し、多数の計算機により共有されるものが存在する。この記憶装置においては、多くの場合、ホストが認識する論理的記憶装置を実際にデータを記憶する物理記憶装置の適当な領域に割り当てることが行われる。このような記憶装置を利用する場合、人間がすべてを把握することは困難であり、このような記憶装置を含む計算機システム側に何かしらのサポート機能が存在しなければ文献 1 に述べられている問題点を把握することすら困難となる。また、問題点を把握することができたとしても、計算機システム側にデータの移動機能が存在しない場合には、記憶装置上のデータの再配置のためにデータのバックアップとリストアが必要となり、その処理に多大な労力を必要とする。

【 0 0 0 9 】

文献 2 で述べられている技術では、ボリュームマネージャレイヤにおけるオブジェクトの稼動状況によるデータ再配置案を作成する機能を実現しているが、記憶装置から更に高いアクセス性能を得ようとする場合には DBMS レイヤにおけるデータの特徴を考慮して配置を決定する必要があるがその点の解決方法については何も述べていない。

## 【 0 0 1 0 】

文献 3, 文献 4 で述べられている技術においては、記憶装置を利用するアプリケーションが利用するデータに関する特徴としてアクセス頻度とシーケンシャルアクセス性程度しか考慮していないため、アプリケーションから見た場合に必ずしも良好な配置が実現できるわけではない。例えば、DBMS では表データとそれに対応する索引データを同時にアクセスすることが多いが、これらのデータを同一の物理記憶装置に配置する可能性がある。この場合、物理記憶装置においてアクセス競合が発生し、記憶装置のアクセス性能が低下する可能性がある。

## 【 0 0 1 1 】

本発明の第一の目的は、DBMS が管理するデータの特性を考慮して記憶装置におけるデータ記憶位置を決定することにより、DBMS に対してより好ましいアクセス性能特性を持つ記憶装置を保持する計算機システムを実現し、DBMS の性能を向上させることである。特に、複数の記憶装置を利用する DB システムにおいて、各記憶装置間へアクセス要求が適切に分散化されるようにし、DBMS の処理性能を向上させる。

## 【 0 0 1 2 】

本発明の第二の目的は、DBMS が稼動している計算機システムにおいて、DBMS の特性を考慮した記憶装置の良好なアクセス性能特性達成を目的としたデータ記憶位置再配置処理を実現することにより、計算機システムの性能に関する管理コストを削減することである。

## 【 0 0 1 3 】

## 【課題を解決するための手段】

DBMS に関する情報を 1 箇所に集約させ、そこでデータの再配置案を決定し、そのデータ再配置案に従うようにデータの移動指示を発行することにより、DBMS に対してより好ましい性能特性を持つデータ配置を実現する。

## 【 0 0 1 4 】

計算機システム内には、複数の DBMS が稼動し、複数の記憶装置が利用されている可能性がある。そのため、DBMS や記憶装置の構成その他の情報を一箇所に集約し、そこで集約した全情報を考慮してデータの再配置案を作成する。

## 【 0 0 1 5 】

DBMSが管理するデータの配置決定方法として、以下のものを採用する。データ更新時に必ず書き込みが実行される更新ログを、他のデータと異なる物理記憶装置に配置して相互干渉しないようにすることによりDBMSに対して良好な性能特性を得ることができる。同時にアクセスされる可能性が極めて高い表データとそれに対応する索引データを異なる物理記憶装置に配置することによりDBMSに対して良好な性能特性を得ることができる。DBMSに関する情報を利用して、データがシーケンシャルにアクセスされる場合のアクセス順序を予測し、その構造を保持するように物理記憶装置に記憶する。

## 【 0 0 1 6 】

現在、計算機のオペレーティングシステム（OS）・データ転送経路中のスイッチ・記憶装置内部においてデータの記憶位置を変更する技術が存在する。データの記憶位置の変更はそれらの技術を用いて実現する。そこで、前記の項目を考慮して決定されたデータ再配置案に従って、データの記憶位置の変更を管理する部分に対してデータ配置変更の指示を発行する。

## 【 0 0 1 7 】

## 【発明の実施の形態】

以下、本発明の実施の形態を説明する。なお、これにより本発明が限定されるものではない。

## ＜第一の実施の形態＞

本実施形態では、DBMSが実行される計算機と記憶装置がスイッチを用いて接続された計算機システムにおいて、データの記憶位置の管理を行う計算機が存在し、そこで計算機システム内のデータの記憶位置の管理を行う。本実施例におけるスイッチは、記憶装置から提供される記憶領域を組み合わせて仮想的な記憶装置を作成する機能を有する。また、記憶装置においても、記憶装置内部でデータの記憶位置を動的に変更する機能を有する。

## 【 0 0 1 8 】

データ記憶位置管理を実施する計算機は、DBMSに関する情報、データの記憶位置のマッピングに関する情報、記憶装置の構成情報を取得し、それらを用い



て好適なデータ再配置案を作成する。スイッチと記憶装置に対して作成したデータ配置を指示し、それらのデータ再配置機能を用いてそのデータ再配置案に従ったデータ配置を実現する。

## 【 0 0 1 9 】

図 1 は、本発明の第一の実施の形態における計算機システムの構成図である。本実施の形態における計算機システムは、DBホスト 8 0、データ位置管理サーバ 8 2、記憶装置 1 0 から構成される。DBホスト 8 0、データ位置管理サーバ 8 2、記憶装置 1 0 はそれぞれが保有するネットワークインターフェイス 7 8 を通してネットワーク 7 9 に接続されている。また、DBホスト 8 0、記憶装置 1 0 はそれぞれが保有する I/O パスインターフェイス 7 0 から I/O パス 7 1 を介して仮想ボリュームスイッチ 7 2 に接続され、これらを通して記憶装置 1 0 と DBホスト 8 0 間のデータ転送を行う。

## 【 0 0 2 0 】

本実施の形態においては、記憶装置 1 0 と DBホスト 8 0 間のデータ転送を行う I/O パス 7 1 とネットワーク 7 9 を異なるものとしているが、例えば i S C S I のような計算機と記憶装置間のデータ転送をネットワーク上で実施する技術も開発されており、本実施の形態においてもこの技術を利用してもよい。このとき、仮想ボリュームスイッチ 7 2 において、I/O パス 7 1 とネットワーク 7 9 との間でデータ転送が可能であれば記憶装置 1 0 もしくは DBホスト 8 0 において I/O パスインターフェイス 7 0 がネットワークインターフェイス 7 8 を兼ねても良い。

## 【 0 0 2 1 】

記憶装置 1 0 は、記憶領域を提供するもので、その記憶領域は記憶領域管理単位であるボリュームを用いて外部に提供し、ボリューム内の部分領域に対するアクセスや管理はブロックを単位として実行する。記憶装置 1 0 は、ネットワークインターフェイス 7 8、I/O パスインターフェイス 7 0、記憶装置制御装置 1 2、ディスクコントローラ 1 6、物理記憶装置 1 8 から構成され、ネットワークインターフェイス 7 8、I/O パスインターフェイス 7 0、記憶装置制御装置 1 2、ディスクコントローラ 1 6 はそれぞれ内部バス 2 0 により接続され、ディス

クコントローラ 1 6 と物理記憶装置 1 8 は物理記憶装置バス 2 2 により接続される。

#### 【 0 0 2 2 】

記憶装置制御装置 1 2 は、CPU 2 4 とメモリ 2 6 を有する。メモリ 2 6 上には、記憶装置におけるキャッシュメモリとして利用するデータキャッシュ 2 8 が割り当てられ、記憶装置を制御するためのプログラムである記憶装置制御プログラム 4 0 が記憶される。また、メモリ 2 6 上には、物理記憶装置 1 8 の稼動情報である物理記憶装置稼動情報 3 2、記憶装置 1 0 が提供するボリュームを物理的に記憶する物理記憶装置 1 8 上の記憶位置の管理情報であるボリューム物理記憶位置管理情報 3 6 を保持する。

#### 【 0 0 2 3 】

図中の記憶装置 1 0 は、複数の物理記憶装置 1 8 を有し、1 つのボリュームに属するデータを複数の物理記憶装置 1 8 に分散配置することが可能である。また、データが記憶される物理記憶装置 1 8 上の位置を動的に変更する機能を有する。このデータ移動指示は、ネットワークインターフェイス 7 8 を通して外部から行うことが可能である。ただし、それらは必須のものではなく、1 つの物理記憶装置の記憶領域をそのままボリュームとして提供する記憶装置 1 0 でも本実施の形態にあてはめることができる。

#### 【 0 0 2 4 】

仮想ボリュームスイッチ 7 2 は、記憶装置 1 0 から提供される記憶領域管理単位であるボリュームの一部領域または全領域を 1 つ以上組み合わせた仮想的なボリュームである仮想ボリュームを DB ホスト 8 0 に提供する機能を有する。仮想ボリュームスイッチ 7 2 は、ネットワークインターフェイス 7 8 を有し、仮想ボリュームスイッチ 7 2 において形成されるボリュームに関する情報である仮想ボリューム情報 7 3 を保持する。

#### 【 0 0 2 5 】

仮想記憶スイッチ 7 2 はホストから行われたアクセス要求を仮想ボリューム情報 7 3 を参照して適切な記憶装置 1 0 の適切な記憶領域へのアクセスへと変換してアクセス処理を実現する。また、ある仮想ボリュームのデータが記憶されるボ

リウムやそのボリューム内の記憶位置を動的に変更する機能を有する。このデータ移動指示は、ネットワークインターフェイス 7 8 を通して外部から行うことが可能である。

## 【 0 0 2 6 】

DBホスト 8 0、データ位置管理サーバ 8 2 においては、それぞれ CPU 8 4、ネットワークインターフェイス 7 8、メモリ 8 8 を有し、メモリ 8 8 上にオペレーティングシステム (OS) 1 0 0 が記憶・実行されている。

## 【 0 0 2 7 】

DBホスト 8 0 は I/O パスインターフェイス 7 0 を有し、仮想ボリュームスイッチ 7 2 により提供されるボリュームに対してアクセスを実行する。OS 1 0 0 内にファイルシステム 1 0 4 と 1 つ以上のボリュームからホストが利用する論理的なボリュームである論理ボリュームを作成するボリュームマネージャ 1 0 2 と、ファイルシステム 1 0 4 やボリュームマネージャ 1 0 2 により、OS 1 0 0 によりアプリケーションに対して提供されるファイルや論理ローボリュームに記憶されたデータの記録位置等を管理するマッピング情報 1 0 6 を有する。

## 【 0 0 2 8 】

OS 1 0 0 が認識する仮想ボリュームやボリュームマネージャ 1 0 2 により提供される論理ボリュームに対して、アプリケーションがそれらのボリュームをファイルと等価なインターフェイスでアクセスするための機構であるローデバイス機構を OS 1 0 0 が有していても良い。図中の構成ではボリュームマネージャ 1 0 2 が存在しているが、本実施の形態においてはボリュームマネージャ 1 0 2 における論理ボリュームの構成を変更することはないので、ボリュームマネージャ 1 0 2 が存在せずにファイルシステムが直接仮想ボリュームスイッチ 7 2 により提供される仮想ボリュームを利用する構成に対しても本実施の形態を当てはめることができる。

## 【 0 0 2 9 】

DBホスト 8 0 のそれぞれのメモリ 8 8 上では DBMS 1 1 0、データ位置管理副プログラム 1 2 0 が記憶・実行され、実行履歴情報 1 2 2 が記憶されている。DBMS 1 1 0 は内部にスキーマ情報 1 1 4 を有している。図中では、DMB

S 1 1 0 が 1 台のホストに 1 つのみ動作しているが、後述するように、D M B S 1 1 0 毎の識別子を用いて管理を行うため、1 台のホストに複数動作していても本実施の形態にあてはめることができる。

#### 【 0 0 3 0 】

データ位置管理サーバ 8 2 のメモリ 8 8 上ではデータ位置管理主プログラム 1 3 0 が記憶・実行され、記憶装置 1 0 内の物理記憶装置 1 8 の稼動情報である記憶装置稼動情報 1 3 2、各記憶装置 1 0 における物理構成やボリュームの物理記憶位置に関する情報である記憶装置構成情報 1 3 4、DBホスト 8 0 上の D B M S 1 1 0 におけるスキーマに関する情報のうちデータ位置管理に必要なデータである D B M S スキーマ情報 1 3 6、D B M S 1 1 0 における D B M S 処理の実行履歴に関する情報である D B M S 実行履歴情報 1 3 8、DBホスト 8 0 における O S 1 0 0 内のマッピング情報 1 0 6 と仮想ボリュームスイッチ 7 2 内の仮想ボリューム情報 7 3 に関する情報を含むデータ記憶位置管理情報 1 4 0 が記憶される。

#### 【 0 0 3 1 】

図中では、データ位置管理サーバ 8 2 は DBホスト 8 0 と異なる計算機となっているが、任意の DBホスト 8 0 がデータ位置管理サーバ 8 2 の役割を兼ねても本実施の形態に当てはめることができる。更に、仮想ボリュームスイッチ 7 2 上や任意の記憶装置 1 0 上にデータ位置管理サーバ 8 2 が提供する機能を実装しても本実施の形態に当てはめることができる。

#### 【 0 0 3 2 】

図 2 は記憶装置 1 0 内に保持されている物理記憶装置稼動情報 3 2 を示す。物理記憶装置稼動情報 3 2 中には、記憶装置 1 0 が提供するボリュームの識別子であるボリューム名 5 0 1 とそのボリューム名 5 0 1 を持つボリュームのデータを保持する物理記憶装置 1 8 の識別子である物理記憶装置名 5 0 2、そしてボリューム名 5 0 1 を持つボリュームが物理記憶装置名 5 0 2 を持つ物理記憶装置 1 8 に記憶しているデータをアクセスするための稼動時間のある時刻からの累積値である累積稼働時間 5 0 3 の組を保持する。

#### 【 0 0 3 3 】

記憶装置 1 0 内の記憶装置制御部 1 2 はディスクコントローラ 1 6 を利用して物理記憶装置 1 8 へのデータアクセスする際の開始時刻と終了時刻を取得し、そのアクセスデータがどのボリュームに対するものかを判断して開始時刻と終了時刻の差分を稼働時間として対応するボリューム名 5 0 1 と物理記憶装置名 5 0 2 を持つデータの組の累積稼働時間 5 0 3 に加算する。この情報は、必ずしも全ての記憶装置 1 0 で取得する必要はない。

## 【 0 0 3 4 】

図 3 は記憶装置 1 0 内に保持されているボリューム物理記憶位置管理情報 3 6 を示す。ボリューム物理記憶位置管理情報 3 6 中には、データの論理アドレス物理記憶装置 1 8 における記憶位置のマッピングを管理するボリューム物理記憶位置メイン情報 5 1 0 と記憶装置 1 0 内でのボリュームに属するデータの物理記憶位置の変更処理の管理情報であるボリュームデータ移動管理情報 5 1 1 が含まれる。

## 【 0 0 3 5 】

ボリューム物理記憶位置メイン情報 5 1 0 中には、ボリューム名 5 0 1 とそのボリューム上のデータ記憶位置であるボリューム論理ブロック番号 5 1 2 とその論理ブロックが記憶されている物理記憶装置 1 8 の識別子である物理記憶装置名 5 0 2 と物理記憶装置 1 8 上の記憶位置である物理ブロック番号 5 1 4 の組のデータが含まれる。ここで、ボリューム名 5 0 1 が “E m p t y” であるエントリ 5 1 5 は特殊なエントリであり、このエントリは記憶装置 1 0 内の物理記憶装置 1 8 の領域のうち、ボリュームに割り当てられていない領域を示し、この領域に対してデータをコピーすることによりデータの物理記憶位置の動的変更機能を実現する。

## 【 0 0 3 6 】

ボリュームデータ移動管理情報 5 1 1 はボリューム名 5 0 1 と、そのボリューム内の記憶位置を変更するデータ範囲を示す移動論理ブロック番号 7 8 2 と、そのデータが新規に記憶される物理記憶装置 1 8 の識別子とその記憶領域を示す移動先物理記憶装置名 7 8 3 と移動先物理ブロック番号 7 8 4、現在のデータコピー元を示すコピーポインタ 7 8 6 とデータの再コピーの必要性を管理する差分管

理情報 7 8 5 の組が含まれる。

【 0 0 3 7 】

差分管理情報 7 8 5 とコピーポインタ 7 8 6 を用いたデータの記憶位置変更処理の概略を以下に示す。差分管理情報 7 8 5 はある一定量の領域毎にデータコピーが必要である「1」または不必要「0」を示すデータを保持する。データの記憶位置変更処理開始時に全ての差分管理情報 7 8 5 のエントリを 1 にセットし、コピーポインタ 7 8 6 を移動元の先頭にセットする。

【 0 0 3 8 】

コピーポインタ 7 8 6 にしたがって差分管理情報 7 8 5 に 1 がセットされている領域を順次移動先にデータをコピーし、コピーポインタ 7 8 6 を更新していく。差分管理情報 7 8 5 で管理される領域をコピーする直前に、その対応するエントリを 0 にセットする。データコピー中に移動領域内のデータに対する更新が行われた場合、それに対応する差分管理情報 7 8 5 のエントリを 1 にセットする。

【 0 0 3 9 】

一度全領域のコピーが完了した段階で差分管理情報 7 8 5 内のエントリが全て 0 になったかを確認し、全て 0 であればボリューム物理記憶位置メイン情報 5 1 0 を更新してデータの記憶位置変更処理は完了する。1 のエントリが残っている場合には、再度それに対応する領域をコピーする処理を前記手順で繰り返す。なお、データ記憶位置の動的変更機能の実現方法は他の方式を用いても良い。この場合には、ボリューム物理記憶位置管理情報 3 6 中にはボリュームデータ移動管理情報 5 1 1 ではなく他のデータ記憶位置の動的変更機能のための管理情報が含まれることになる。

【 0 0 4 0 】

図 4 は DB ホスト 8 0 の OS 1 0 0 内に記憶されているマッピング情報 1 0 6 を示す。マッピング情報 1 0 6 中には、ボリュームローデバース情報 5 2 0、ファイル記憶位置情報 5 3 0 と論理ボリューム構成情報 5 4 0 が含まれる。ボリュームローデバース情報 5 2 0 中には OS 1 0 0 においてローデバースを指定するための識別子であるローデバースパス名 5 2 1 とそのローデバースによりアクセスされる仮想ボリュームあるいは論理ボリュームの識別子であるローデバースボ

リユーム名 5 2 2 の組が含まれる。

【 0 0 4 1 】

ファイル記憶位置情報 5 3 0 中には、OS 1 0 0 においてファイルを指定するための識別子であるファイルパス名 5 3 1 とそのファイル中のデータ位置を指定するブロック番号であるファイルブロック番号 5 3 2 とそれに対応するデータが記憶されている仮想ボリユームもしくは論理ボリユームの識別子であるファイル配置ボリユーム名 5 3 3 とそのボリユーム上のデータ記憶位置であるファイル配置ボリユームブロック番号 5 3 4 の組が含まれる。

【 0 0 4 2 】

論理ボリユーム構成情報 5 4 0 中にはボリユームマネージャ 1 0 2 により提供される論理ボリユームの識別子である論理ボリユーム名 5 4 1 とその論理ボリユーム上のデータの位置を示す論理ボリユーム論理ブロック番号 5 4 2 とその論理ブロックが記憶されている仮想ボリユームの識別子である仮想ボリユーム名 5 4 3 と仮想ボリユーム上の記憶位置である仮想ボリユームブロック番号 5 4 4 の組が含まれる。

【 0 0 4 3 】

図 5 は DBMS 1 1 0 内に記憶されているその内部で定義・管理しているデータその他の管理情報であるスキーマ情報 1 1 4 を示す。スキーマ情報 1 1 4 には、表のデータ構造や制約条件等の定義情報を保持する表定義情報 5 5 1、索引のデータ構造や対象である表等の定義情報を保持する索引定義情報 5 5 2、利用するログに関する情報であるログ情報 5 5 3、利用する一時表領域に関する情報である一時表領域情報 5 5 4、管理しているデータのデータ記憶位置の管理情報であるデータ記憶位置情報 5 5 5、キャッシュの構成に関する情報であるキャッシュ構成情報 5 5 6 とデータをアクセスする際の並列度に関する情報である最大アクセス並列度情報 5 5 7 を含む。

【 0 0 4 4 】

データ記憶位置情報 5 5 5 中には、表、索引、ログ、一時表領域等のデータ構造の識別子であるデータ構造名 5 6 1 とそのデータを記憶するファイルまたはローデバイスの識別子であるデータファイルパス名 5 6 2 とその中の記憶位置であ

るファイルブロック番号 5 6 3 の組が含まれる。

【 0 0 4 5 】

キャッシュ構成情報 5 5 6 は DBMS 1 1 0 が三種類のキャッシュ管理のグループを定義し、そのグループに対してキャッシュを割り当てている場合を示す。キャッシュ構成情報 5 5 6 中には、グループ名 5 6 5 とグループ中のデータ構造のデータをホスト上にキャッシュする際の最大データサイズであるキャッシュサイズ 5 6 6 とそのグループに所属するデータ構造の識別子の所属データ構造名 5 6 7 の組が含まれる。最大アクセス並列度情報 5 5 7 には、データ構造名 5 6 1 とそのデータ構造にアクセスする際の一般的な場合の最大並列度に関する情報である最大アクセス並列度 5 6 9 の組が含まれる。

【 0 0 4 6 】

図 6 は DB ホスト 8 0 のメモリ 8 8 上に記憶されている実行履歴情報 1 2 2 を示す。実行履歴情報 1 2 2 中には、DBMS 1 1 0 で実行されたクエリ 5 7 0 の履歴が記憶されている。この情報は、DBMS 1 1 0 が作成する。または DBMS のフロントエンドプログラムがこの情報を作成する。この場合には、DBMS フロントエンドプログラムが存在する計算機に実行履歴情報 1 2 2 が記憶されることになる。

【 0 0 4 7 】

図 7 は仮想ボリュームスイッチ 7 2 が保持する仮想ボリューム情報 7 3 を示す。仮想ボリューム情報 7 3 は仮想ボリューム記憶位置情報 7 9 0 と仮想ボリュームデータ移動管理情報 7 9 1 を含む。仮想ボリューム記憶位置情報 7 9 0 中には、仮想ボリュームスイッチ 7 2 が DB ホスト 8 0 に提供する仮想ボリュームの識別子である仮想ボリューム名 5 4 3 とその仮想ボリューム上のデータの記憶位置を示す仮想ボリュームブロック番号 5 4 4 とそのブロックが記憶されている記憶装置 1 0 の識別子である記憶装置名 5 8 3 とそのボリュームの識別子であるボリューム名 5 1 1 とボリューム上の記憶位置であるボリューム論理ブロック番号 5 1 2 の組が含まれる。

【 0 0 4 8 】

仮想ボリューム名 5 4 3 が “E m p t y” であるエントリ 5 8 5 は特殊なエン



トリであり、このエントリに含まれる記憶装置 1 0 上の領域は D B ホスト 8 0 に対して仮想ボリュームとして提供されていない領域を示す。これらの領域に対して仮想ボリュームスイッチ 7 2 はデータの移動を行うことができる。

## 【 0 0 4 9 】

仮想ボリュームデータ移動管理情報 7 9 1 は仮想ボリューム名 5 4 3 と、そのボリューム内の記憶位置を変更するデータ範囲を示す移動仮想ボリュームブロック番号 7 9 3 と、そのデータが新規に記憶される記憶装置 1 0 の識別子とその記憶領域を示す移動先記憶装置名 7 9 4 と移動先ボリューム名 7 9 5 と移動先論理ブロック番号 7 9 6、現在のデータコピー元を示すコピーポインタ 7 8 6 とデータの再コピーの必要性を管理する差分管理情報 7 8 5 の組が含まれる。

## 【 0 0 5 0 】

図 3 中のボリュームデータ移動管理情報 5 1 1 で説明したのと同様の方式によりデータの記憶位置の動的変更機能を実現できる。データ記憶位置の動的変更機能の実現方法は他の方式を用いても良い。この場合には、仮想ボリューム情報 7 3 中には仮想ボリュームデータ移動管理情報 7 9 1 ではなく他のデータ記憶位置の動的変更機能のための管理情報が含まれることになる。

## 【 0 0 5 1 】

図 8 はデータ位置管理サーバ 8 2 上に記憶される記憶装置稼働情報 1 3 2 を示す。記憶装置稼働情報 1 3 2 中には、記憶装置 1 0 の識別子である記憶装置名 5 8 3、記憶装置 1 0 が提供するボリュームの識別子であるボリューム名 5 0 1、記憶装置 1 0 に存在する物理記憶装置 1 8 の識別子である物理記憶装置名 5 0 2、記憶装置名 5 8 3 とボリューム名 5 0 1、物理記憶装置名 5 0 2 により特定される領域の前回取得時の累積稼働時間 5 0 3 の値である旧累積稼働時間 5 9 3 とある一定時間内の動作時間の割合を示す稼働率 5 9 4 の組と、稼働率計算のために前回累積稼働時間を取得した時刻である前回累積稼働時間取得時刻 5 9 5 を含む。

## 【 0 0 5 2 】

記憶装置 1 0 は物理記憶装置稼働情報 3 2 を外部に提供する機構を有し、それを利用して記憶位置管理主プログラム 1 3 0 は記憶装置 1 0 で取得・記憶されて

いる物理記憶装置稼働情報 3 2 をネットワーク 7 9 を通して一定間隔で取得し、取得した累積稼働時間 5 0 3 と旧累積稼働時間 5 9 3、前回累積稼働時間取得時刻 5 9 5 と現データ取得時刻を用いて前回累積稼働時間取得時刻 5 9 5 と現データ取得時刻間の稼働率 5 9 4 を計算・記憶する。その後、取得した累積稼働時間 5 0 3 を旧累積稼働時間 5 9 3 に、現データ取得時刻を前回累積稼働時間取得時刻 5 9 5 に記憶する。

## 【 0 0 5 3 】

なお、全ての記憶装置 1 0 で物理記憶装置稼働情報 3 2 を保持しているとは限らない。その場合には物理記憶装置稼働情報 3 2 を保持している記憶装置 1 0 についてのみ記憶装置稼働情報 1 3 2 のエントリに含める。また、全ての記憶装置 1 0 で物理記憶装置稼働情報 3 2 を保持していない場合には記憶装置稼働情報 1 3 2 を保持しなくてもよい。

## 【 0 0 5 4 】

図 9 はデータ位置管理サーバ 8 2 上に記憶される記憶装置構成情報 1 3 4 を示す。記憶装置構成情報 1 3 4 中には、記憶装置 1 0 の識別子である記憶装置名 5 8 3 と、記憶装置 1 0 がデータ記憶位置の動的変更機能を有しているかいないかの情報である移動機能情報 6 0 1 と記憶装置 1 0 が保持しているデータキャッシュの容量であるデータキャッシュ容量 6 0 2、記憶装置名 5 8 3 を持つ記憶装置 1 0 におけるボリューム物理記憶位置メイン情報 5 1 0 を保持する記憶装置ボリューム物理記憶位置管理情報 6 0 3 の組を保持する。

## 【 0 0 5 5 】

記憶装置 1 0 はボリューム物理記憶位置メイン情報 5 1 0 とデータキャッシュ 2 8 のサイズに関する情報を外部に提供する機構を有し、記憶装置構成情報 1 3 4 を作成するため、データ記憶位置管理主プログラム 1 3 0 は記憶装置 1 0 からボリューム物理記憶位置情報 3 6 とデータキャッシュ 2 8 のサイズに関する情報をネットワーク 7 9 を通して取得する。記憶装置 1 0 はデータキャッシュ 2 8 のサイズは必ずしも外部に提供する機能を有する必要はなく、その場合には、データキャッシュ容量 6 0 2 の対応部分はデータ無効を記憶しておく。

## 【 0 0 5 6 】

一方、ボリューム物理記憶位置メイン情報 5 1 0 に関しては、記憶装置 1 0 が外部に提供する機能を有さなくても構わない場合は、物理記憶装置 1 8 を 1 つしか保持せずにそれをそのまま 1 つのボリュームとして提供する等、記憶装置 1 0 が提供するボリュームがどのように物理記憶装置 1 8 上に記憶されるかが固定され、かつ、そのマッピングをあらかじめデータ記憶位置管理主プログラム 1 3 0 が理解している場合である。このとき、データ記憶位置管理主プログラム 1 3 0 はこのルールに従って記憶装置ボリューム物理記憶位置管理情報 6 0 3 の内容を設定する。このルールは設定ファイル等を用いて管理者がデータ記憶位置管理主プログラム 1 3 0 に与える。

## 【 0 0 5 7 】

図 1 0 はデータ位置管理サーバ 8 2 上に記憶される DBMS スキーマ情報 1 3 6 を示す。DBMS スキーマ情報 1 3 6 は、DBMS データ構造情報 6 2 1、DBMS データ記憶位置情報 6 2 2、DBMS パーティション化表・索引情報 6 2 3、DBMS 索引定義情報 6 2 4、DBMS キャッシュ構成情報 6 2 5、DBMS ホスト情報 6 2 6 を含む。

## 【 0 0 5 8 】

DBMS データ構造情報 6 2 1 は DBMS 1 1 0 で定義されているデータ構造に関する情報で、DBMS 1 1 0 の識別子である DBMS 名 6 3 1、DBMS 1 1 0 内の表・索引・ログ・一時表領域等のデータ構造の識別子であるデータ構造名 5 6 1、データ構造の種別を表すデータ構造種別 6 4 0、そのデータ構造をアクセスする際の最大並列度に関する情報である最大アクセス並列度 5 6 9 の組を保持する。このとき、データ構造によっては最大アクセス並列度 5 6 9 の値を持たない。

## 【 0 0 5 9 】

DBMS データ記憶位置情報 6 2 2 は DBMS 名 6 3 1 とその DBMS におけるデータ記憶位置管理情報 5 5 5 であるデータ記憶位置管理情報 6 3 8 の組を保持する。DBMS パーティション化表・索引情報 6 2 3 は、1 つの表や索引をある属性値により幾つかのグループに分割したデータ構造を管理する情報で、パーティション化されたデータ構造が所属する DBMS 1 1 0 の識別子である DBM

S名631と分割化される前のデータ構造の識別子であるパーティション元データ構造名643と分割後のデータ構造の識別子であるデータ構造名561とその分割条件を保持するパーティション化方法644の組を保持する。今後、パーティション化されたデータ構造に関しては、特に断らない限り単純にデータ構造と呼ぶ場合にはパーティション化後のものを指すものとする。

#### 【0060】

DBMS索引定義情報624には、DBMS名631、索引の識別子である索引名635、その索引のデータ形式を示す索引タイプ636、その索引がどの表のどの属性に対するものかを示す対応表情報637の組を保持する。DBMSキャッシュ構成情報625は、DBMS110のキャッシュに関する情報であり、DBMS名631とDBMS110におけるキャッシュ構成情報556の組を保持する。DBMSホスト情報626は、DBMS名631を持つDBMS110がどのホスト上で実行されているかを管理するもので、DBMS名631とDBMS実行ホストの識別子であるホスト名651の組を保持する。

#### 【0061】

DBMSスキーマ情報136中のDBMSホスト情報626以外は、データ記憶位置管理主プログラム130がDBMS110が管理しているスキーマ情報114の中から必要な情報を取得して作成するものである。DBMS100のスキーマ情報114は、ネットワーク79を通してデータ位置管理主プログラム130が直接、あるいは、データ位置管理副プログラム120を介して、SQL等のデータ検索言語を用いてビューとして公開されている情報を取得するか、または、専用の機構を用いて取得する。DBMSホスト情報626はシステム構成情報であり、管理者が設定する。

#### 【0062】

図11はデータ位置管理サーバ82上に記憶されるDBMS実行履歴情報138を示す。DBMS実行履歴情報138には、DBMS110の識別子のDBMS名631と各々のDBMS110で実行されたクエリ570の履歴が保持される。これは、データ位置管理主プログラム130がネットワーク79を通してDBホスト80内に記憶されている実行履歴情報122をデータ位置管理副プログ

ラム 1 2 0 を利用して収集し、それを記憶したものである。

【 0 0 6 3 】

また、前述のように実行履歴情報 1 2 2 が DBMS フロントエンドプログラムが実行される計算機上に記憶される可能性も存在する。この場合には、DBMS フロントエンドプログラムが実行される計算機から実行履歴情報 1 2 2 をデータ位置管理サーバ 8 2 へ転送する手段を設け、転送された実行履歴情報 1 2 2 をデータ位置管理主プログラム 1 3 0 が DBMS 実行履歴情報 1 3 8 として記憶する。なお、本実施の形態においては、全ての DBMS 1 1 0 から実行履歴情報 1 2 2 を収集する必要はなく、DBMS 実行履歴情報 1 3 8 は存在しなくてもよい。

【 0 0 6 4 】

図 1 2 はデータ位置管理サーバ 8 2 上に記憶されるデータ記憶位置管理情報 1 4 0 を示す。データ記憶位置管理情報 1 4 0 には、ホストマッピング情報 6 5 0 と仮想ボリューム記憶位置管理情報 7 9 0 が含まれる。ホストマッピング情報 6 5 0 には、ホストの識別子であるホスト名 6 5 1 とそのホストにおけるマッピング情報 1 0 6 の組が保持される。これは、データ位置管理主プログラム 1 3 0 がネットワーク 7 9 を通して、DB ホスト 8 0 の OS 1 0 0 が保持しているマッピング情報 1 0 6 を、データ位置管理副プログラム 1 2 0 を利用して収集し、それを記憶したものである。

【 0 0 6 5 】

データ位置管理副プログラム 1 2 0 は、OS 1 0 0 が提供している管理コマンドや情報提供機構、参照可能な管理データの直接解析等によりマッピング情報 1 0 6 を取得する。仮想ボリュームスイッチ 7 2 は外部に仮想ボリューム記憶位置管理情報 7 9 0 を提供する機構を有し、データ位置管理主プログラム 1 3 0 はネットワーク 7 9 を介して仮想ボリュームスイッチ 7 2 から仮想ボリューム記憶位置管理情報 7 9 0 を取得する。

【 0 0 6 6 】

図 1 3 はデータ位置管理主プログラム 1 3 0 によるデータ再配置処理の処理フローを示す。ここで、処理開始は管理者の指示によることとする。後述するように、複数の異なった種類のデータ配置解析・データ再配置案作成処理を実行可能

であり、処理すべき種類の指定をして処理を開始する。また、処理にパラメータが必要な場合は併せてそれを管理者が指示をするものとする。本実施の形態においては、データの記憶位置の動的変更機能は仮想ボリュームスイッチ 7 2 と記憶装置 1 0 が保持する。ただし、記憶装置 1 0 はデータの記憶位置の動的変更機能を有さなくてもよい。

## 【 0 0 6 7 】

ステップ 2 0 0 1 でデータ再配置処理を開始する。このとき、データ配置解析・データ再配置案作成処理として何を実行するか指定する。また、必要であればパラメータを指定する。

## 【 0 0 6 8 】

ステップ 2 0 0 2 でデータ再配置処理に必要な情報を収集し、記憶装置稼働情報 1 3 2、記憶装置構成情報 1 3 4、DBMS スキーマ情報 1 3 6、DBMS 実行履歴情報 1 3 8、データ記憶位置管理情報 1 4 0 として記憶する。なお、このデータ収集は、ステップ 2 0 0 1 の処理開始とは無関係にあらかじめ実行しておくこともできる。この場合には、情報を取得した時点から現在まで情報に変更がないかどうかをこのステップで確認する。

## 【 0 0 6 9 】

ステップ 2 0 0 3 では、ワーク領域を確保し、その初期化を行う。ワーク領域としては、図 1 4 に示すデータ再配置ワーク情報 6 7 0 と作成した図 1 5 に示す移動プラン情報 7 5 0 を利用する。データ再配置ワーク情報 6 7 0 と移動プラン情報 7 5 0 の詳細とその初期データ作成方法は後述する。

## 【 0 0 7 0 】

ステップ 2 0 0 4 でデータ配置の解析・再配置案の作成処理を実行する。後述するように、データ配置の解析・再配置案作成処理は複数の観点による異なったものが存在し、このステップではステップ 2 0 0 1 で指定された処理を実行する。またステップ 2 0 0 1 でパラメータを受け取った場合には、それを実行する処理に与える。

## 【 0 0 7 1 】

ステップ 2 0 0 5 ではステップ 2 0 0 4 のデータ再配置案作成処理が成功した

かどうか確認する。成功した場合にはステップ 2 0 0 7 に進む。失敗した場合にはステップ 2 0 0 6 に進み、管理者にデータ再配置案作成が失敗したことを通知し、ステップ 2 0 1 0 に進み処理を完了する。

#### 【 0 0 7 2 】

ステップ 2 0 0 7 では、ステップ 2 0 0 4 で作成されたデータ再配置案を管理者に提示する。この提示を受けて管理者はデータ再配置案に問題がないか判断する。

ステップ 2 0 0 8 では、データの再配置を続行するか否かを管理者から指示を受ける。続行する場合にはステップ 2 0 0 9 に進み、そうでない場合にはステップ 2 0 1 0 に進み処理を完了する。

#### 【 0 0 7 3 】

ステップ 2 0 0 9 では、ステップ 2 0 0 4 で作成されたデータの再配置案を基にデータの再配置指示を仮想ボリュームスイッチ 7 2、あるいは、記憶装置 1 0 に対して発行する。仮想ボリュームスイッチ 7 2 と記憶装置 1 0 はネットワーク 7 9 を通してデータのデータ再配置指示を受ける機能を有し、これを利用して指示が出される。

#### 【 0 0 7 4 】

この指示の形式としては、仮想ボリュームスイッチ 7 2 に対しては、それが提供する仮想ボリュームのあるデータ領域を、記憶装置 1 0 のボリューム内のデータ領域を指定してそこへの移動を指示するものとなり、記憶装置 1 0 に対しては、それが提供するボリュームのあるデータ領域を、その記憶装置 1 0 内の物理記憶装置 1 8 のデータ領域を指定してそこへの移動を指示するものとなる。この指示に従って、仮想ボリュームスイッチ 7 2、あるいは、記憶装置 1 0 はデータの再配置処理を実行する。

#### 【 0 0 7 5 】

ステップ 2 0 1 0 でデータ再配置処理は完了である。

#### 【 0 0 7 6 】

この処理フローは、管理者の指示で処理を開始し、作成されたデータ再配置案に問題点がないかステップ 2 0 0 7 と 2 0 0 8 で管理者が判断している。この管

理者による確認を省略し、タイマを用いて設定した処理開始時刻にデータ配置解析・再配置案作成処理を開始することによりデータ再配置処理の自動化も可能である。

## 【 0 0 7 7 】

図 1 4 は図 1 3 のステップ 2 0 0 3 において作成する情報であるデータ再配置ワーク情報 6 7 0 を示す。データ再配置ワーク情報 6 7 0 中には、仮想ボリューム物理記憶位置情報 6 8 0 とデータ構造仮想ボリューム内位置情報 6 9 0 が含まれる。

## 【 0 0 7 8 】

仮想ボリューム物理記憶位置情報 6 8 0 は仮想ボリュームスイッチ 7 2 が提供する仮想ボリュームのデータがどの記憶装置 1 0 内のどの物理記憶装置 1 8 のどの位置に記憶されているかに関する情報であり、仮想ボリューム名 5 4 3、仮想ボリュームブロック番号 5 4 4、記憶装置名 5 8 3、ボリューム名 5 0 1、物理記憶装置名 5 0 2、物理ブロック番号 5 1 4 の組で表される。仮想ボリューム物理記憶位置情報 6 8 0 では、データ記憶位置管理情報 1 4 0 中の仮想ボリューム記憶位置管理情報と、記憶装置構成情報 1 3 4 中の記憶装置名 5 8 3 と記憶装置ボリューム物理記憶位置情報 6 0 3 を参照し、記憶装置 1 0 が提供するボリュームに関して対応する部分をまとめることにより初期データが作成される。

## 【 0 0 7 9 】

仮想ボリューム名 5 4 3 が “E m p t y” であるエントリ群 6 8 1 は記憶装置 1 0 内の物理記憶装置 1 8 の記憶領域のうち、データ再配置のためにデータを移動することが可能な領域の集合を表し、データ再配置案作成時に、この領域の中から適切なデータの移動先を発見する。このうち、ボリューム名 5 0 1 が有効値のエントリは、仮想ボリュームスイッチ 7 2 におけるデータ記憶位置の動的変更に利用可能な領域であり、その利用には特に制約はない。一方のボリューム名 5 0 1 が無効値のエントリは、記憶装置 1 0 におけるデータ記憶位置の動的変更に利用可能な領域であり、その領域が属する記憶装置 1 0 内に記憶されているデータの記憶位置変更に関してのみ利用可能である。

## 【 0 0 8 0 】



データ構造仮想ボリューム内位置情報 6 9 0 は、DBMS 1 1 0 が保持しているデータ構造が仮想ボリュームスイッチ 7 2 が提供する仮想ボリュームのどこに記憶されているかを示した情報であり、DBMS 名 6 3 1、データ構造名 5 6 1、データファイルパス名 5 6 2、ファイルブロック番号 5 6 3、仮想ボリューム名 5 4 3、仮想ボリュームブロック番号 5 4 4 の組を保持する。この情報では、DBMS スキーマ情報 1 3 6 中の DBMS データ記憶位置情報 6 2 2 と DBMS ホスト情報 6 2 6 とデータ記憶位置管理情報 1 4 0 中のホストマッピング情報 6 5 0 を参照し、ファイル（ローデバイス）パス、論理（仮想）ボリュームに関して対応する部分をまとめることにより初期データが作成される。

## 【 0 0 8 1 】

図 1 5 は図 1 3 のステップ 2 0 0 4 で実行されるデータ配置解析・データ再配置案作成処理により作成されるデータ移動案を格納する移動プラン情報 7 5 0 を示す。移動プラン情報 7 5 0 は、仮想ボリュームスイッチ 7 2 に対するデータ移動指示を記憶する仮想ボリューム移動プラン情報 7 5 1 と記憶装置 1 0 に対するデータ移動指示を記憶する物理記憶位置移動プラン情報 7 5 2 を含む。これら情報に関しては、何もデータを持たないように初期化する。

## 【 0 0 8 2 】

仮想ボリューム移動プラン情報 7 5 1 には、移動指示の実行順序を示す移動順序 7 6 1、移動するデータを持つ仮想ボリュームとそのデータ領域を示す移動仮想ボリューム名 7 6 2 と移動仮想ボリュームブロック番号 7 6 3、そのデータの移動先の記憶装置、ボリューム、ボリューム内の記憶領域を示す移動先記憶装置名 7 6 4 と移動先ボリューム名 7 6 5 と移動先ボリューム論理ブロック番号 7 6 6 の組を保持する。

## 【 0 0 8 3 】

物理記憶位置移動プラン情報 7 5 2 には、移動指示の実行順序を示す移動順序 7 6 1、移動するデータを持つ記憶装置 1 0 とボリュームとそのデータ領域を示す移動記憶装置名 7 6 7 と移動ボリューム名 7 6 8 と移動ボリューム論理ブロック番号 7 6 9、そのデータの移動先の物理記憶装置とその記憶領域を示す移動先物理記憶装置名 7 7 1 と移動先物理ブロック番号 7 7 2 の組を保持する。なお、物

理記憶位置移動プラン情報 7 5 2 に関しては、全ての記憶装置 1 0 が記憶装置内の物理記憶位置の動的変更機能を有さない場合には、この情報を保持する必要はない。

## 【 0 0 8 4 】

図 1 3 のステップ 2 0 0 4 で実行されるデータ配置解析・データ再配置案作成処理について説明する。前述のように、この処理には幾つかの種類が存在する。全ての処理に共通するのは逐次的にデータ再配置するためのデータ移動案を作成することである。そのため、データ移動の順番には意味があり、移動プラン情報 7 5 0 中の移動順序 7 6 1 にその順番を保持し、その順序どおりにデータ移動を行うことによりデータの再配置を実施する。また、逐次処理のため、移動後のデータ配置をもとに次のデータの移動方法を決定する必要がある。そこで、データ移動案を作成するたびにデータ再配置ワーク情報 6 7 0 をデータ移動後の配置に更新する。

## 【 0 0 8 5 】

データ再配置案作成時のデータ移動案の作成は以下のように行う。移動したいデータ量以上の連続した移動可能領域をデータ再配置ワーク情報 6 7 0 中の情報から抽出し、その中の領域を適当に選択し、設定条件や後述する制約を満たすかどうか確認をする。もし、それらを満たす場合にはそこを移動先として設定する。それらを満たさない場合には他の領域を選択し、再度それらを満たすかどうか確認をする。以下、設定条件と制約を満たす領域を検出するか、全ての移動したいデータ量以上の連続した移動可能領域が設定条件や制約を満たさないことを確認するまで処理を繰り返す。もし、全ての領域で設定条件や制約を満たさない場合にはデータ移動案の作成に失敗として終了する。

## 【 0 0 8 6 】

このときに重要なのは移動後のデータ配置において、問題となる配置を行わないことである。特に R D B M S においては、特定のデータに関してはアクセスが同時に行われる可能性が高く、それらを異なる物理記憶装置 1 8 上に配置する必要がある。

## 【 0 0 8 7 】

そこで、以下で説明する全てのデータの移動案を作成する場合には、移動するデータに含まれるデータ構造と、移動先に含まれるデータ構造を調べ、ログとその他のデータ、一時表領域とその他のデータ、表データとそれに対して作成された木構造の索引データがデータの移動後に同じ物理記憶装置 1 8 に配置されるかどうかを確認し、配置される場合には、その配置案は利用不可能と判断する。

## 【 0 0 8 8 】

なお、あるデータ構造がどの記憶装置 1 0 のどのボリューム上の領域に、あるいは、どの物理記憶装置 1 8 の領域に記憶されているか、また逆に、ある物理記憶装置 1 8 上や記憶装置 1 0 のボリュームの領域に記憶されるデータがどのデータ構造に対応するかは、データ再配置ワーク情報 6 7 0 中の仮想ボリューム物理記憶位置情報 6 8 0 とデータ構造仮想ボリューム内位置情報 6 9 0 を仮想ボリュームに関して対応する部分を組み合わせることにより把握可能である。

## 【 0 0 8 9 】

図 1 6 に第 1 のデータ配置解析・データ再配置案作成処理である、記憶装置稼動情報 1 3 2 を基にした同時アクセス実行データ構造を分離するためのデータ再配置案作成処理の処理フローを示す。本処理においては、物理記憶装置 1 8 の稼動率が閾値を超えたものはディスクネック状態にあると判断してそれを解消するデータの移動案を作成する。

## 【 0 0 9 0 】

前述のように、記憶装置稼動情報 1 3 2 は全ての記憶装置 1 0 中の物理記憶装置 1 8 に関する稼動情報を必ずしも含む訳ではない。稼動情報が存在しない物理記憶装置 1 8 に関しては、本処理におけるデータ再配置案作成の対象外とし、それらが存在しないものとして扱う。本処理は、実測値に基づいて問題点を把握し、それを解決する方法を見つけるため、より精度の高いデータ再配置案を作成するものであり、ボトルネックの自動解決方式としてデータ移動の自動化機能に組み入れても効果的に働く。

## 【 0 0 9 1 】

ステップ 2 1 0 1 で処理を開始する。本処理を開始するにあたっては、どの期間の稼動率を参照するかを指定する。

## 【 0 0 9 2 】

ステップ 2 1 0 2 では、物理記憶装置 1 8 の識別子と指定期間における物理記憶装置 1 8 の稼働率の組を記憶するワーク領域を取得し、記憶装置稼働情報 1 3 2 を参照してその情報を設定し、物理記憶装置 1 8 の稼働率で降順にソートする。記憶装置稼働情報 1 3 2 中では、同じ物理記憶装置 1 8 中に記憶されているデータであっても異なるボリュームのものは分離して稼働率を取得しているため、それらの総和として物理記憶装置 1 8 の稼働率を求める必要がある。ステップ 2 1 0 3 では、ステップ 2 1 0 2 のソート結果をもとに物理記憶装置 1 8 の稼働率が閾値を超えているもののリストである過負荷確認リストを作成する。このリスト中のエントリに関しても稼働率が降順になるような順序を保つようにする。

## 【 0 0 9 3 】

ステップ 2 1 0 4 では、過負荷確認リスト中にエントリが存在するか確認する。エントリが存在しない場合には、もう過負荷状態の物理記憶装置 1 8 が存在しないものとしてステップ 2 1 0 5 に進みデータ再配置案作成処理成功として処理を終了する。エントリが存在する場合には、ステップ 2 1 0 6 に進む。

## 【 0 0 9 4 】

ステップ 2 1 0 6 では、過負荷確認リスト中の最も物理記憶装置 1 8 の稼働率が高いものを再配置対象の物理記憶装置 1 8 として選択する。ステップ 2 1 0 7 では、再配置対象となった物理記憶装置 1 8 内部のボリュームとその稼働率のリストを記憶装置稼働情報 1 3 2 を参照して作成し、稼働率で降順にソートする。

## 【 0 0 9 5 】

ステップ 2 1 0 8 では、リスト中のあるボリュームの稼働率があらかじめ定められた閾値を超過しているかどうか確認する。全てのボリュームの稼働率が閾値を超えていない場合には、ステップ 2 1 1 3 に進み、あるボリュームの稼働率はその閾値を超えている場合には、ステップ 2 1 0 9 に進む。

## 【 0 0 9 6 】

ステップ 2 1 0 9 においては、稼働率が閾値を超えているボリュームに関して、確認対象の物理記憶装置 1 8 中に同時にアクセスされる可能性があるデータの組、すなわち、ログとその他のデータ、一時表領域とその他のデータ、表データ

とそれに対して作成された木構造の索引データがあるそのボリューム内部に記憶されているかどうか検出する処理を行う。ステップ 2 1 1 0 では、ステップ 2 1 0 9 における結果を確認し、同時アクセスデータ構造の組が存在する場合にはステップ 2 1 1 1 に進む。同時アクセスデータ構造の組が存在しない場合には、ステップ 2 1 1 2 に進む。

## 【 0 0 9 7 】

ステップ 2 1 1 1 においては、同時アクセスデータ構造の組に属するデータを異なる物理記憶装置 1 8 に記憶するためのデータ移動案を作成し、ステップ 2 1 1 4 に進む。

## 【 0 0 9 8 】

ステップ 2 1 1 2 においては、現在確認対象となっているのボリューム内のデータを論理ブロック番号に従って 2 分割し、その片方を他の物理記憶装置 1 8 へ移動するデータ移動案を作成し、ステップ 2 1 1 4 に進む。

## 【 0 0 9 9 】

ステップ 2 1 1 3 においては、現在確認対象になっている物理記憶装置 1 8 の稼働率が閾値を下回るまで、稼働率が高いボリュームから順に、その物理記憶装置 1 8 に記憶されているボリュームを構成するデータ全体を他の物理記憶装置 1 8 に移動するデータ移動案を作成し、ステップ 2 1 1 4 に進む。

## 【 0 1 0 0 】

ステップ 2 1 1 1, 2 1 1 2, 2 1 1 3 のデータ移動先を検出する際に、移動後の移動先の記憶装置の稼働率を予測する。物理記憶装置 1 8 毎の性能差が既知の場合にはその補正を行った移動データを含む記憶装置 1 8 上のボリュームの稼働率分、未知の場合には補正を行わない移動データを含む記憶装置 1 8 上のボリュームの稼働率分、データ移動により移動先の物理記憶装置 1 8 の稼働率が上昇すると考え、加算後の値が閾値を越えないような場所へのデータの移動案を作成する。稼働率の加算分に関して、移動データ量の比率を考慮しても良いが、ここではデータ中のアクセスの偏りを考慮して移動データに全てのアクセスが集中したと考えた判断を行う。

## 【 0 1 0 1 】

ステップ 2 1 1 4 では、データの移動案の作成に成功したかどうかを確認し、失敗した場合にはステップ 2 1 1 7 に進みデータの再配置案作成処理失敗として処理を終了する。成功した場合にはステップ 2 1 1 5 に進む。

#### 【 0 1 0 2 】

ステップ 2 1 1 5 では作成したデータ移動案を移動プラン情報 7 5 0 に追加、ステップ 2 1 1 6 に進む。ステップ 2 1 1 6 ではデータ再配置ワーク情報 6 7 0 を作成したデータ移動案に従って修正し、移動先記憶装置 1 8 のステップ 2 1 0 2 で作成した物理記憶装置 1 8 毎の稼働情報情報の値を前述の移動後の稼働率判断値に修正する。その後、現在の確認対象の物理記憶装置 1 8 を過負荷確認リストから削除し、ステップ 2 1 0 4 に戻り次の確認を行う。

#### 【 0 1 0 3 】

次に第 2 のデータ配置解析・データ再配置案作成処理である、DBMS 実行履歴情報 1 3 8 を基にした同時アクセス実行データ構造を分離するためのデータ再配置案作成処理を示す。本処理においては、クエリの実行履歴から同時にアクセスされるデータの組を取得し、それらを異なる物理記憶装置 1 8 に配置するデータ再配置案を作成する。前述のように、全ての DBMS 1 1 0 について実行履歴を取得できるわけではない。本処理実行時に実行履歴が存在しない DBMS 1 1 0 が利用するデータに関してはデータ再配置の対象外とする。

#### 【 0 1 0 4 】

本処理においては、図 1 7 に示すクエリ実行時同時アクセスデータカウント情報 7 0 0 を利用する。クエリ実行時同時アクセスデータカウント情報 7 0 0 は、DBMS 名 6 3 1、同時にアクセスされる可能性のあるデータ構造のデータ構造名 5 6 1 の組を示すデータ構造名 A 7 0 1 とデータ構造名 B 7 0 2、そして、DBMS 実行履歴情報 1 3 8 の解析によりそのデータ構造の組がアクセスされたと判断された回数であるカウント値 7 0 3 の組で表される。この組はカウント値 7 0 3 の値でソートしておく。

#### 【 0 1 0 5 】

クエリ実行時同時アクセスデータカウント情報 7 0 0 は DBMS 実行履歴情報 1 3 8 から作成する。最初にクエリ実行時同時アクセスデータカウント情報 7 0

0のエントリを全消去する。DBMS100において定型処理が行われる場合には、まず、その型により分類し、その型の処理が何回実行されたかを確認する。

## 【0106】

続いてDBMS100から型毎のクエリ実行プランを取得する。そのクエリ実行プランにより示される処理手順から同時にアクセスされるデータ構造の組を判別する。そして、クエリ実行時同時アクセスデータカウント情報700中のDBMS名631・データ構造名A701・データ構造名B702を参照し、既に対応するデータ構造の組が存在している場合には先に求めたその型の処理回数をカウント値703に加算する。既に対応するデータ構造の組が存在していない場合には、新たにエントリを追加してカウント値703を先に求めたその型の処理回数にセットする。

## 【0107】

DBMS100において非定型処理が行われる場合には、1つ1つの実行されたクエリに関してクエリ実行プランを取得し、そのクエリ実行プランにより示される処理手順から同時にアクセスされるデータ構造の組を判別する。そして、クエリ実行時同時アクセスデータカウント情報700中のDBMS名631・データ構造名A701・データ構造名B702を参照し、既に対応するデータ構造の組が存在している場合にはカウント値703に1を加算する。既に対応するデータ構造の組が存在していない場合には、新たにエントリを追加してカウント値703に1をセットする。

## 【0108】

クエリ実行プランから同時にアクセスされる可能性があるデータ構造の判別は以下のように行う。まず、木構造の索引に対するアクセスが実施される場合には、その木構造索引データと、その索引が対象とする表データが同時にアクセスされると判断する。また、データの更新処理や挿入処理が行われる場合には、ログとその他のデータが同時にアクセスされると判断する。以下はDBMS110の特性に依存するが、例えば、クエリ実行プラン作成時にネストループジョイン処理を多段に渡り実行する計画を作成し、それらの多段に渡る処理を同時に実行するRDBMSが存在する。このRDBMSを利用する場合にはその多段に渡るネ

ストループジョイン処理で利用する表データとその表に対する木構造の索引データは同時にアクセスされると判断できる。

#### 【0109】

このように、クエリ実行計画による同時アクセスデータの判断に関しては、DBMS 110の処理特性を把握して判断する必要があるが、ここでは、対象とするDBMS 110の種類を絞りデータ位置管理主プログラム130がDBMS 110特有の同時アクセスデータ構造の組を把握できる機能を有することを仮定する。

#### 【0110】

図18にDBMS実行履歴情報138を基にした同時アクセス実行データ構造を分離するためのデータ再配置案作成処理処理フローを示す。ステップ2201で処理を開始する。ステップ2202で実行履歴から同時にアクセスされるデータ構造の組とその実行頻度に関する情報である、前述のクエリ実行同時アクセスデータ構造カウント情報700を作成する。ステップ2203において、カウント値703の全エントリの総和に対してカウント値703の値が一定割合以上のデータ構造とその所属するDBMS 110の組を求め、それらを確認リストとして記憶する。

#### 【0111】

ステップ2204でステップ2203で求めた確認リスト中に含まれるデータ構造の組に関して、それらを異なる物理記憶装置18に記憶するデータ再配置案を作成し、ステップ2205に進む。なお、ステップ2204の処理に関しては、図19を用いて後で説明する。ステップ2205では、ステップ2204においてデータ再配置案の作成に成功したかどうかを確認し、成功した場合にはステップ2206に進みデータ再配置案作成処理成功として処理を終了し、失敗した場合にはステップ2207に進みデータ再配置案作成処理失敗として処理を終了する。

#### 【0112】

図19に、指定されたデータ構造とそのデータ構造と同時にアクセスされる可能性が高いデータ構造の組を分離するデータ再配置案を作成する処理のフローを



示す。本処理を開始するときには、データ構造名と物理記憶装置 1 8 から分離するデータ構造名の組のリストである確認リストを与える。

## 【0 1 1 3】

ステップ 2 3 0 1 で処理を開始する。ステップ 2 3 0 3 で確認リスト中にエントリが存在するか確認し、存在しない場合にはステップ 2 3 0 4 に進みデータ再配置案作成処理成功として処理を終了する。存在する場合にはステップ 2 3 0 5 に進む。

## 【0 1 1 4】

ステップ 2 3 0 5 においては、確認リストから 1 つ確認対象データ構造名とその所属 DBMS 名の組とその分離データ構造名とその所属 DBMS 名の組の組を取得し、ステップ 2 3 0 6 に進む。

## 【0 1 1 5】

ステップ 2 3 0 6 においては、確認対象データ構造とその分離するデータ構造が同一の物理記憶装置上に記憶されているかどうかの確認を行う。前述のように、この確認はデータ再配置ワーク情報 6 7 0 を参照することにより可能である。両データ構造が全て異なる物理記憶装置上に存在する場合にはステップ 2 3 1 2 に進み、ある物理記憶装置上に両データ構造が存在する場合にはステップ 2 3 0 7 に進む。

## 【0 1 1 6】

ステップ 2 3 0 7 においては、同一の物理記憶装置上に両データ構造が存在する部分に関してそれを分離するデータ移動案を作成する。ステップ 2 3 0 8 においては、そのデータ移動案作成が成功したかどうか確認し、成功した場合にはステップ 2 3 1 0 に進み、失敗した場合にはステップ 2 3 0 9 に進みデータ再配置案作成処理失敗として処理を終了する。

## 【0 1 1 7】

ステップ 2 3 1 0 においては、作成されたデータ移動案を移動プラン情報 7 5 0 に記憶する。ステップ 2 3 1 1 においては、作成されたデータ移動案に従ってデータ再配置ワーク情報 6 7 0 を更新し、ステップ 2 3 1 2 に進む。

## 【0 1 1 8】

ステップ2312においては、確認リストから現在確認対象となっているデータ構造の組のエントリを削除し、ステップ2303に進む。

## 【0119】

図20に、第3のデータ配置解析・データ再配置案作成処理である、データ構造の定義を基にした同時アクセス実行データ構造を分離するためのデータ再配置案作成処理の処理フローを示す。本処理においては、同時にアクセスされる可能性が高い、ログとその他のデータ、一時表領域とその他のデータ、表データとそれに対して作成された木構造の索引データが同一物理記憶装置18上に記憶されている部分が存在しないか確認をし、そのような部分が存在する場合にはそれを解決するデータ再配置案を作成する。

## 【0120】

ステップ2401で処理を開始する。ステップ2402では、DBMSデータ構造情報621を参照して全てのログであるデータ構造名561とそれを利用するDBMS110のDBMS名631の組を取得する。そして、そのデータ構造名とログ以外のデータを分離するデータ構造とする確認リストを作成し、ステップ2403に進む。

## 【0121】

ステップ2403ではステップ2402で作成した確認リストを用いてステップ2301から開始されるデータ構造分離のためのデータ再配置案作成処理を実行する。ステップ2404ではステップ2403におけるデータ再配置案作成処理が成功したか確認をし、成功した場合にはステップ2405に進む。失敗した場合にはステップ2412に進みデータ再配置案作成処理失敗として処理を終了する。

## 【0122】

ステップ2405では、DBMSデータ構造情報621を参照して全ての一時表領域であるデータ構造名561とそれを利用するDBMS110のDBMS名631の組を取得する。そして、そのデータ構造名と一時表領域以外のデータを分離するデータ構造とする確認リストを作成し、ステップ2406に進む。

## 【0123】

ステップ 2 4 0 6 ではステップ 2 4 0 5 で作成した確認リストを用いてステップ 2 3 0 1 から開始されるデータ構造分離のためのデータ再配置案作成処理を実行する。ステップ 2 4 0 7 ではステップ 2 4 0 6 におけるデータ再配置案作成処理が成功したか確認をし、成功した場合にはステップ 2 4 0 8 に進む。失敗した場合にはステップ 2 4 1 2 に進みデータ再配置案作成処理失敗として処理を終了する。

## 【 0 1 2 4 】

ステップ 2 4 0 8 では、DBMS 索引定義情報 6 2 4 を参照して全ての木構造索引の索引名 6 3 5 とそれに対応する表のデータ構造名を対応表情報 6 3 7 から取得する。そして、索引名 6 3 5 と対応する表のデータ構造名とそれらを保持する DBMS 1 1 0 の DBMS 名 6 3 1 を組とする確認リストを作成し、ステップ 2 4 0 9 に進む。

## 【 0 1 2 5 】

ステップ 2 4 0 9 ではステップ 2 4 0 8 で作成した確認リストを用いてステップ 2 3 0 1 から開始されるデータ構造分離のためのデータ再配置案作成処理を実行する。ステップ 2 4 1 0 ではステップ 2 4 0 9 におけるデータ再配置案作成処理が成功したか確認をし、成功した場合にはステップ 2 4 1 1 に進み、データ再配置案作成処理成功として処理を終了する。失敗した場合にはステップ 2 4 1 2 に進みデータ再配置案作成処理失敗として処理を終了する。

## 【 0 1 2 6 】

図 2 1 に第 4 のデータ配置解析・データ再配置案作成処理である、特定の表や索引の同一データ構造に対するアクセス並列度を考慮したデータ再配置案作成処理の処理フローを示す。この処理は、ランダムアクセス実行時の処理の並列度を考慮してディスクネックの軽減を図るためにデータの再配置を行うものである。この処理を実行する際には、データ再配置の確認対象とするデータ構造を DBMS 名 6 3 1 とデータ構造名 5 6 1 の組として指定する。

## 【 0 1 2 7 】

ステップ 2 5 0 1 で処理を開始する。ステップ 2 5 0 2 において、指定されたデータ構造の物理記憶装置上に割り当てられた記憶領域利用総量を求める。この

値は、指定データ構造がデータファイル上に割り当てられた容量と等しいため、DBMSデータ記憶位置情報622中のデータ記憶位置を参照してその容量を求めればよい。

## 【0128】

ステップ2503においては、DBMSデータ構造情報621を参照して指定データ構造における最大アクセス並列度569を取得する。ステップ2504において、ステップ2502で求めた指定データ構造の記憶領域利用総量をステップ2503で求めた最大アクセス並列度569で割った値を、指定データ構造の1つの物理記憶装置18上への割り当てを許可する最大量として求める。この制約により、特定の物理記憶装置18に偏ることなく最大アクセス並列度569以上の台数の物理記憶装置18に指定データ構造が分散して記憶されることになり、最大アクセス並列度569による並列度でランダムアクセスが実行されてもディスクネックになりにくい状況となる。なお、割り当て許可最大量の値は、実際のアクセス特性を考慮してこの方法で求めた値から更に増減させても構わない。

## 【0129】

ステップ2505において、指定データ構造のデータがステップ2504で求めた最大量を超えて1つの物理記憶装置18上に割り当てられているものが存在するかデータ再配置ワーク情報670を用いて確認し、そのようなものが存在しない場合にはステップ2509に進み、データ再配置案作成処理成功として処理を終了する。存在する場合にはステップ2506に進む。

## 【0130】

ステップ2506においては、ステップ2504で求めた最大量を超えて1つの物理記憶装置18上に割り当てられている部分を解消するデータ移動案を作成する。このとき、移動案作成に考慮するデータ移動量は指定データ構造の現在の物理記憶装置18上への割り当て量のステップ2504で求めた最大量からの超過分以上である必要がある。また、移動先物理記憶装置18においても、移動後にステップ2504で求めた最大量を超えないようにする必要がある。

## 【0131】

ステップ2507においては、ステップ2506のデータ移動案作成処理が成

功したか確認をする。成功した場合にはステップ 2 5 0 8 に進む。失敗した場合にはステップ 2 5 1 0 に進み、データ再配置案作成処理失敗として処理を終了する。

#### 【 0 1 3 2 】

ステップ 2 5 0 8 においては作成したデータ移動案を移動プラン情報 7 5 0 に記憶し、ステップ 2 5 0 9 に進みデータ再配置案作成処理成功として処理を終了する。

#### 【 0 1 3 3 】

図 2 2 に第 5 のデータ配置解析・データ再配置案作成処理である、特定の表データに対するシーケンシャルアクセス時のディスクネックを解消するデータ再配置案作成処理の処理フローを示す。この処理を実行する際には、データ再配置の確認対象とする表を DBMS 名 6 3 1 とデータ構造名 5 6 1 の組として指定する。

#### 【 0 1 3 4 】

DBMS 1 1 0 毎により、シーケンシャルアクセス方法が定まっている。そこで、対象とする DBMS 1 1 0 の種類を絞り、あらかじめデータ位置管理主プログラム 1 3 0 が DBMS 1 1 0 におけるシーケンシャルアクセス方法を把握し、それらに対する最適化を行う。本実施の形態の DBMS 1 1 0 におけるシーケンシャルアクセス方法は以下の方法に従うものとする。あるデータ構造のデータをシーケンシャルアクセスする場合に、データ構造が記憶されているデータファイル名 5 6 2 とファイルブロック番号 5 6 3 を昇順にソートしその順序でアクセスを実行する。

#### 【 0 1 3 5 】

その他にシーケンシャルアクセス方法の決定方法としては、データファイルを管理する内部通番とファイルブロック番号 5 6 3 の組を昇順にソートした順番にアクセスする方法等が存在し、それらを利用したシーケンシャルアクセス方法の判断を実施してもよい。

#### 【 0 1 3 6 】

また、並列にシーケンシャルアクセスを実行する場合に、その領域の分割法は

並列にアクセスしない場合のシーケンシャルにアクセスする順番を並列度に合わせて等分に分割するものとする。

## 【 0 1 3 7 】

この並列アクセスによる分割後の 1 つのアクセス領域を全て同一の物理記憶装置 1 8 上に配置するのは必ずしも現実的ではない。そこで、分割後のアクセス領域がある一定量以上連続にまとまって 1 つの物理記憶装置上に記憶されていればよいと判断する。ただし、どのような場合でも連続してアクセスされることがなく、分割後のアクセス領域が異なるものに分類されるものに関しては、並列シーケンシャルアクセス時にアクセスがぶつかる可能性があるため、異なる物理記憶装置 1 8 に記憶するという指針を設けて、これに沿うようなデータ配置を作成することによりシーケンシャルアクセスの性能を高める。

## 【 0 1 3 8 】

ステップ 2 6 0 1 で処理を開始する。ステップ 2 6 0 2 において、指定された表の物理記憶装置上に割り当てられた記憶領域利用総量を求める。この値は、指定データ構造がデータファイル上に割り当てられた容量と等しいため、DBMS データ記憶位置情報 6 2 2 中のデータ記憶位置を参照してその容量を求めればよい。ステップ 2 6 0 3 においては、DBMS データ構造情報 6 2 1 を参照して指定データ構造における最大アクセス並列度 5 6 9 を取得する。

## 【 0 1 3 9 】

ステップ 2 6 0 4 において、ステップ 2 6 0 2 で求めた指定表の記憶領域利用総量をステップ 2 6 0 3 で求めた最大アクセス並列度 5 6 9 で割った量が、並列アクセス時にシーケンシャルにアクセスされる 1 つの領域のデータ量である。そこで、データ位置管理主プログラム 1 3 0 が把握している前述のシーケンシャルアクセス先の決定方法に基づいて DBMS データ記憶位置情報 6 2 2 中の指定データ構造のデータファイルにおける記憶先を求め、それらのアクセス方法を前述のように予測し、その結果をもとに最大アクセス並列度 5 6 9 の並列アクセスが実行されると仮定した前述のデータ分割指針を作成する。

## 【 0 1 4 0 】

ステップ 2 6 0 5 において、データ再配置ワーク情報 6 7 0 を参照しながら、

指定データ構造はステップ 2 6 0 4 で作成したデータ分割指針に沿ったデータ配置が物理記憶装置 1 8 上で行われているか確認し、そうであればステップ 2 6 0 9 に進み、データ再配置案作成処理成功として処理を終了する。そうでない場合にはステップ 2 6 0 6 に進む。

## 【 0 1 4 1 】

ステップ 2 6 0 6 においては、物理記憶装置 1 8 上において、ステップ 2 6 0 4 で求めたデータ分割指針に従うデータ配置を求める。このとき、データがある一定値以下の領域に細分化されている場合には、連続した空き領域を探し、そこにアクセス構造を保つようにデータを移動するデータ移動案を作成する。また、最大アクセス並列度 5 6 9 の並列アクセスにより異なるアクセス領域に分離されるデータが同じ物理記憶装置 1 8 上に配置されないようなデータ移動案を作成する。

## 【 0 1 4 2 】

ステップ 2 6 0 7 においては、ステップ 2 6 0 6 のデータ移動案作成処理が成功したか確認をする。成功した場合にはステップ 2 6 0 8 に進み、失敗した場合にはステップ 2 6 1 0 に進み、データ再配置案作成処理失敗として処理を終了する。

## 【 0 1 4 3 】

ステップ 2 6 0 8 においては、作成したデータ移動案を移動プラン情報 7 5 0 に記憶し、ステップ 2 6 0 9 に進みデータ再配置案作成処理成功として処理を終了する。

## 【 0 1 4 4 】

次に第 6 のデータ配置解析・データ再配置案作成処理である、特定のデータ構造に対する記憶装置 1 0 におけるキャッシュ効果を考慮したデータ再配置案作成処理の説明を行う。この処理を実行する際には、データ再配置の確認対象とするデータ構造として DBMS 名 6 3 1 とデータ構造名 5 6 1 を指定する。また、処理実行時に、記憶装置におけるキャッシュ効果が存在するかどうかを明示的に指定してもよい。前述のように、記憶装置構成情報 1 3 4 において、すべての記憶装置 1 0 がデータキャッシュ容量 6 0 2 に有効な値を保持しているわけではない

。そのため、データキャッシュ容量 6 0 2 が無効値である記憶装置 1 0 は本処理の対象外とする。

## 【 0 1 4 5 】

本処理においては、指定データ構造には記憶装置 1 0 のキャッシュ効果が存在するかどうかを判断する。まず、データ構造の単位データ量あたりのホストキャッシュにおける平均的キャッシュ利用量を計算する。その結果を用いて十分な量のホストキャッシュが利用可能かを判断する。十分な量のホストキャッシュを利用可能な場合には、アクセス頻度が低いデータに対してのみ記憶装置からデータを読み出すことになり、記憶装置におけるキャッシュ効果は極めて低いものとなる。

## 【 0 1 4 6 】

その確認処理のフローを図 2 3 に示す。ステップ 2 8 0 1 で処理を開始する。ステップ 2 8 0 2 で DBMS キャッシュ構成情報 6 2 5 を参照して指定データ構造が属するキャッシュグループを求め、それから DBMS データ記憶位置情報を参照してそのキャッシュグループに属するデータ構造の記憶のために割り当てられた領域量の総和を求める。

## 【 0 1 4 7 】

ステップ 2 8 0 3 において、DBMS キャッシュ構成情報 6 2 5 から指定データ構造が属するキャッシュグループに割り当てられたキャッシュサイズ 5 6 6 を求め、それとステップ 2 8 0 2 で求めた領域量の総和から指定データ構造のホストにおける単位データ量あたりの平均キャッシュ利用量を求め、それをキャッシュ効果判断閾値と比較する。単位データ量あたりの平均キャッシュ利用量が閾値以上の場合にはステップ 2 8 0 4 に進み、指定データ構造は記憶装置におけるキャッシュ効果がないと判断し、ステップ 2 8 0 6 に進み処理を終了する。閾値未満の場合にはステップ 2 8 0 5 に進み、指定データ構造は記憶装置におけるキャッシュ効果があると判断し、ステップ 2 8 0 6 に進み処理を終了する。

## 【 0 1 4 8 】

本処理において、記憶装置 1 0 におけるデータ構造のキャッシュ利用量を判断する。このとき、実際のデータのキャッシュ利用量はアクセスパターンに依存す



るが、ここでは平均的なケースを考え、データ構造のキャッシュ利用量は、その記憶装置 1 0 に割り当てられたデータ量に比例するとする。つまり、記憶装置構成情報 1 3 4 から記憶装置ボリューム物理記憶位置管理情報 6 0 3 を参照して記憶装置 1 0 における全データ記憶容量を求め、また、データキャッシュ容量 6 0 2 からデータキャッシュ容量を求める。これらの値から単位データ容量あたりのデータキャッシュ量が求まりこの値を元に判断する。また、記憶装置 1 0 において、提供するボリュームをいくつかにグループ化し、それぞれで単位容量あたりのキャッシュ利用量を変化させる制御を行うことが可能である。この場合には、ボリューム等の記憶領域毎にデータ構造のキャッシュ利用量が増加することになる。

## 【 0 1 4 9 】

図 2 4 に特定のデータ構造に対する記憶装置 1 0 におけるキャッシュ効果を考慮したデータ再配置案作成処理の処理フローを示す。ステップ 2 7 0 1 で処理を開始する。ステップ 2 7 0 2 において、データ再配置案の作成対象として指定されたデータ構造に対して、処理開始時にキャッシュ効果があるかないかが明示的に指定されたか確認する。明示的に指定されていない場合にはステップ 2 7 0 3 に進み、明示的に指定された場合にはステップ 2 7 0 4 に進む。

## 【 0 1 5 0 】

ステップ 2 7 0 3 において、前述した指定データ構造の記憶装置 1 0 におけるキャッシュ効果が存在するかどうか確認処理を実行する。

## 【 0 1 5 1 】

ステップ 2 7 0 4 において、指定データ構造は記憶装置 1 0 においてキャッシュ効果があるかないかを確認する。キャッシュ効果があると判断された場合にはステップ 2 7 0 5 に進み、ないと判断された場合にはステップ 2 7 0 6 へ進む。

## 【 0 1 5 2 】

ステップ 2 7 0 5 においては、指定データ構造のデータを十分な量のキャッシュを利用可能な装置、ボリューム等の記憶領域に移動するデータ移動案を作成する。ここで、十分な量のキャッシュを利用可能な記憶領域とは、原則として単位容量あたりのキャッシュ利用量の大きな記憶領域を意味するが、以下の意味も併

せ持つ。指定データ構造のホストキャッシュの利用可能量が大きなものである場合、アクセス頻度が高いものはホストキャッシュに留まることになる。そのため、記憶装置 1 0 におけるキャッシュ利用可能容量は、ホストキャッシュにおける利用可能量に比べて、ある一定比率以上の容量がないとその効果は小さなものと考えられる。そこで、データ構造の単位データ量あたりのホストキャッシュにおける平均的キャッシュ利用量と記憶領域における単位容量あたりの平均的キャッシュ利用量の割合を計算し、その値と判断閾値を比較して記憶領域の方がよりキャッシュを多く利用できると判断される場合でないと十分な量のキャッシュが利用可能であると判断しない。本処理終了後、ステップ 2 7 0 7 に進む。

## 【 0 1 5 3 】

ステップ 2 7 0 6 においては、指定データ構造のデータをキャッシュ利用量を少量に抑えられる領域にデータを移動する、つまり、記憶装置における単位容量あたりの平均的キャッシュ利用量が小さな装置、ボリューム等の領域へデータを移動するデータの移動案を作成し、ステップ 2 7 0 7 に進む。

## 【 0 1 5 4 】

ステップ 2 7 0 7 においては、データ移動案の作成に成功したかどうかを確認する。成功した場合にはステップ 2 7 0 8 に進み、失敗した場合にはステップ 2 7 1 0 に進み、データ再配置案作成処理失敗として処理を終了する。

## 【 0 1 5 5 】

ステップ 2 7 0 8 においては作成したデータ移動案を移動プラン情報 7 5 0 に記憶し、ステップ 2 7 0 9 に進みデータ再配置案作成処理成功として処理を終了する。

## &lt; 第二の実施の形態 &gt;

本実施形態では、DBMS が実行される計算機と記憶装置が接続された計算機システムにおいて、データの記憶位置の管理を行う計算機が存在し、そこで計算機システム内のデータの記憶位置の管理を行う。計算機上で実行される OS 中のファイルシステムやボリュームマネージャは、動的にデータの記憶位置を変更する機能を有する。また、記憶装置においても、記憶装置内部でデータの記憶位置を動的に変更する機能を有する。

## 【0156】

データ記憶位置管理を実施する計算機は、DBMSに関する情報、データの記憶位置のマッピングに関する情報、記憶装置の構成情報を取得し、それらを用いて好適なデータ再配置案を作成する。ファイルシステム、ボリュームマネージャ、記憶装置に対して作成したデータ配置を指示し、それらのデータ再配置機能を用いてそのデータ再配置案に従ったデータ配置を実現する。

## 【0157】

図25は、本発明の第二の実施の形態における計算機システムの構成図である。図示されたように、本発明の第二の実施の形態は本発明の第一の実施の形態と以下の点が異なる。

## 【0158】

仮想ボリュームスイッチ72がI/Oパススイッチ72bに変更され、ネットワーク79とは未接続になる。DBホスト80bで実行されるOS100が有するボリュームマネージャ102がボリュームマネージャ102bに、ファイルシステム104がファイルシステム104bに変更され、OS100が保持するマッピング情報106がマッピング情報106bへとその内容が若干変更される。データ位置管理サーバ82内に記憶されるデータ記憶位置管理情報140がデータ記憶位置管理情報140内のホストマッピング情報650の内容を若干変更したホストマッピング情報650bに変更される。

## 【0159】

I/Oパススイッチ72bは仮想ボリュームスイッチ72と比較して経路制御の機能のみを保持する。本実施の形態においては、記憶装置10とDBホスト80b間のデータ転送を行うI/Oパス71とネットワーク79を異なるものとしているが、例えばiSCSIのような計算機と記憶装置間のデータ転送をネットワーク上で実施する技術も開発されおり、本実施の形態においてもこの技術を利用してもよい。このとき、記憶装置10とDBホスト80bにおいてI/Oパスインターフェイス70が省かれ、計算機システム内からI/Oパス71とI/Oパススイッチ72bが省かれる構成となる。

## 【0160】

ボリュームマネージャ 1 0 2 b はボリュームマネージャ 1 0 2 と比べて、ボリュームマネージャ 1 0 2 b が提供する論理ボリュームの指定領域に記憶されているデータを、指定した記憶装置 1 0 が提供するボリュームの指定した記憶領域に移動する機能を有する。直接的にこの機能を有していない場合でも、管理領域の動的変更機能と管理領域内管理単位毎のデータ移動機能を有するものに関してはそれらの機能の組み合わせで実現できる。このデータ移動機能に関しては、ホスト上で実行される管理コマンドにより処理が実施される。

## 【 0 1 6 1 】

ファイルシステム 1 0 4 b は、ファイルシステム 1 0 4 と比べて、ファイルシステム 1 0 4 b が管理する領域の中でデータが記憶されていない領域に対してファイルデータの一部をそこに移動する機能を有する。そのデータ移動指示方法は、移動するファイルとそのデータ領域と、移動先領域を指定するものとする。このデータ移動機能に関しては、ホスト上で実行される管理コマンドにより処理が実施される。

## 【 0 1 6 2 】

ファイルのデータ記憶位置の動的変更機能は例えばファイルシステムのデフラグ機能として実現されており、その技術を用いてデータの移動先領域を指定できるようにすることにより前記データ移動機能を実現できる。データの移動先となることができる領域は、マッピング情報 1 0 6 b を参照することにより把握することができる。

## 【 0 1 6 3 】

なお、本実施の形態においては、ボリュームマネージャ 1 0 2 b とファイルシステム 1 0 4 b のどちらか一方がデータ移動機能を有していれば良い。また、ファイルシステム 1 0 4 b がデータ移動機能を有している場合には、ボリュームマネージャ 1 0 2 b が存在しなくても本実施の形態に当てはめることができる。

## 【 0 1 6 4 】

図 2 6 は O S 1 0 0 内に記憶されるマッピング情報 1 0 6 b を示す。図 4 のマッピング情報 1 0 6 からの変更点の概略は、マッピング情報 1 0 6 においては仮想ボリュームスイッチ 7 2 が提供する仮想ボリュームを利用していた部分をマッ

ピング情報 1 0 6 b では記憶装置 1 0 が提供するボリュームに変更された部分と、ファイル記憶位置情報 5 3 0 では保持していなかったファイルシステム I D 5 3 5 と空き領域の管理情報がファイル記憶位置情報 5 3 0 b に追加されたことである。

## 【 0 1 6 5 】

マッピング情報 1 0 6 b 中には、ボリュームローデバース情報 5 2 0 b、ファイル記憶位置情報 5 3 0 b と論理ボリューム構成情報 5 4 0 b が含まれる。ボリュームローデバース情報 5 2 0 b 中には O S 1 0 0 においてローデバースを指定するための識別子であるローデバースパス名 5 2 1 と、ローデバースとして記憶装置 1 0 が提供するボリュームを利用する場合の記憶装置 1 0 の識別子である記憶装置名 5 8 3、そのローデバースによりアクセスされるボリュームあるいは論理ボリュームの識別子であるローデバースボリューム名 5 2 2 b の組が含まれる。

## 【 0 1 6 6 】

ファイル記憶位置情報 5 4 0 b 中には、O S 1 0 0 においてファイルを指定するための識別子であるファイルパス名 5 3 1 とそのファイルが存在するファイルシステムのホスト内識別子であるファイルシステム I D 5 3 5 とそのファイル中のデータ位置を指定するブロック番号であるファイルブロック番号 5 3 2 とそれに対応するデータが記憶されている記憶装置 1 0 が提供するボリュームもしくは論理ボリュームの識別子であるファイル配置ボリューム名 5 3 3 b と記憶装置 1 0 が提供するボリュームを利用する場合の記憶装置名 5 8 3 とそのボリューム上のデータ記憶位置であるファイル配置ボリュームブロック番号 5 3 4 の組が含まれる。ファイルパス名 5 3 1 が “E m p t y” であるエントリ 5 3 6 は特殊なエントリであり、これはファイルシステム内における空き領域を示すデータである。この領域に対してデータの移動を行うことができる。

## 【 0 1 6 7 】

論理ボリューム構成情報 5 4 0 b 中にはボリュームマネージャ 1 0 2 b により提供される論理ボリュームの識別子である論理ボリューム名 5 4 1 とその論理ボリューム上のデータの位置を示す論理ボリュームブロック番号 5 4 2 とその論理

ブロックが記憶されている記憶装置10が提供するボリュームの識別子であるボリューム名501とそれを提供する記憶装置名583とボリューム上の記憶位置であるボリューム論理ブロック番号512の組が含まれる。

## 【0168】

図27はデータ位置管理サーバ82上に記憶されるホストマッピング情報650bを示す。図12のホストマッピング情報650からの変更点は、各ホストのマッピング情報106を保持するマッピング情報652が各ホストのマッピング情報106bを保持するマッピング情報652bになったことである。

## 【0169】

データ再配置処理において、データ移動機能を保持する部分が仮想ボリュームスイッチ72からボリュームマネージャ102b、ファイルシステム104bに変更されたことによる変更点は以下のようなものである。

## 【0170】

ボリュームマネージャ102b、ファイルシステム104bにおいては、DBホスト80b上において管理コマンドを実施することによりデータの移動を実施する。そこで、データ位置管理主プログラム130がネットワーク79を通してDBホスト80b上で実行されているデータ位置管理副プログラム120にボリュームマネージャ102b、ファイルシステム104bにおいてデータの移動を実施する管理コマンドを実行する指示を出し、それに従ってデータ位置管理副プログラム120が管理コマンドを実施することによりデータの移動を実施する。

## 【0171】

なお、ボリュームマネージャ102bにおいてデータ移動先として利用する領域は、データ位置管理主プログラム130移動指示を発行したときにはボリュームマネージャ102bの管理下に存在しない可能性がある。この場合には、データ位置管理副プログラム120は、データ移動の管理コマンドを実行する前に移動先領域をボリュームマネージャ102bの管理下に収めるための管理コマンドを実行する。また、データ移動処理が終了後、他のDBホスト80上のボリュームマネージャ102bがデータ移動元の領域を新たなデータ移動先として利用するため、データ移動元領域の開放のための管理コマンドを実施する。

## 【 0 1 7 2 】

また、データ再配置案作成処理においては、データ再配置案を作成する際に利用するワーク領域であるデータ再配置ワーク情報 6 7 0 がデータ再配置ワーク情報 6 7 0 b に、移動プラン情報 7 5 0 が移動プラン情報 7 5 0 b に変更される。

## 【 0 1 7 3 】

図 2 8 にデータ再配置ワーク情報 6 7 0 b を示す。データ再配置ワーク情報 6 7 0 b 中には、ワーク用記憶装置ボリューム記憶位置情報 6 8 2 とワーク用空き領域情報 6 8 3 とデータ構造仮想ボリューム内位置情報 6 9 0 b が含まれる。図 1 4 のデータ再配置ワーク情報 6 7 0 と比べて、仮想ボリューム物理記憶位置情報 6 8 0 はワーク用記憶装置ボリューム記憶位置情報 6 8 2 とワーク用空き領域情報 6 8 3 に分離・変更され、データ構造仮想ボリューム内位置情報 6 9 0 はデータ構造仮想ボリューム内位置情報 6 9 0 b に変更される。

## 【 0 1 7 4 】

ワーク用記憶装置ボリューム記憶位置情報 6 8 2 は記憶装置 1 0 が提供するボリュームとその物理記憶装置 1 8 における記憶位置の一覧であり、記憶装置名 5 8 3、ボリューム名 5 0 1、ボリューム論理ブロック番号 5 1 2、物理記憶装置名 5 0 2、物理ブロック番号 5 1 4 の組を保持する。このデータは、記憶装置構成情報 1 3 4 を参照して初期化する。

## 【 0 1 7 5 】

ワーク用空き領域情報 6 8 3 は、データ再配置案を作成する際にデータの移動先となりうる場所を管理するものであり、ホスト名 6 3 1、ファイルシステム ID 5 3 5、論理ボリューム名 5 4 1、論理ボリュームブロック番号 5 4 2、記憶装置名 5 8 3、ボリューム名 5 0 1、ボリューム論理ブロック番号 5 1 2、物理記憶装置名 5 0 2、物理ブロック番号 5 1 4 の組のデータを保持する。このデータはワーク用空き領域情報 6 8 3 は、ホストマッピング情報 6 5 0 b と記憶装置構成情報 1 3 4 を参照して初期化する。このうち、ホスト名 6 3 1 とファイルシステム ID 5 3 5 に有効値を有するエントリはそれらにより識別されるファイルシステム 1 0 4 b におけるデータ移動先となりうる領域を示し、ホストマッピング情報 6 5 0 b 中のファイル記憶位置情報 5 3 0 b からその領域を把握する。こ

の領域に関しては、そのファイルシステム 1 0 4 b 内に存在するデータの移動先として利用可能である。

#### 【 0 1 7 6 】

ホスト名 6 3 1 とファイルシステム I D 5 3 5 に無効値を有するがボリューム名 5 0 1 に有効値を有するエントリは、まだどのホストからも利用されていない記憶装置 1 0 が提供しているボリューム内の記憶領域を示し、記憶装置構成情報 1 3 4 から把握される記憶装置 1 0 が提供している全領域からホストマッピング情報 6 5 0 b から把握される利用領域を除いた領域として把握する。この領域に関しては、ボリュームマネージャ 1 0 2 b におけるデータ移動先として利用可能である。ボリューム名 5 0 1 に無効値を有するエントリは記憶装置 1 0 内のデータ移動先として利用可能な領域を示し、記憶装置構成情報 1 3 4 から把握可能である。

#### 【 0 1 7 7 】

データ構造仮想ボリューム内位置情報 6 9 0 b は DBMS 1 1 0 が保持しているデータ構造が記憶装置 1 0 が提供するボリュームのどこに記憶されているかを示した情報であり、ホスト名 6 5 1、DBMS 名 6 3 1、データ構造名 5 6 1、データファイルパス名 5 6 2、ファイルシステム I D 5 3 5、ファイルブロック番号 5 6 3、記憶装置名 5 8 3、ボリューム名 5 0 1、ボリューム論理ブロック番号 5 1 2 の組を保持する。この情報は、DBMS スキーマ情報 1 3 6 中の DBMS データ記憶位置情報 6 2 2 と DBMS ホスト情報 6 2 6 とホストマッピング情報 6 5 0 b を参照し、ファイル（ローデバイス）パス、（論理）ボリュームに関して対応する部分をまとめることにより初期データを作成する。

#### 【 0 1 7 8 】

図 2 9 に移動プラン情報 7 5 0 b を示す。移動プラン情報 7 5 0 b は、ボリュームマネージャ 1 0 2 b に対するデータ移動指示を記憶する論理ボリューム移動プラン情報 7 5 3 とファイルシステム 1 0 4 b に対するデータ移動指示を記憶するファイルブロック移動プラン情報 7 5 4 と記憶装置 1 0 に対するデータ移動指示を記憶する物理記憶位置移動プラン情報 7 5 2 を含む。これら情報に関しては、何もデータを持たないように初期化する。図 1 5 の移動プラン情報 7 5 0 と比



較して、移動プラン情報 750b は仮想ボリューム移動プラン情報 752 が削除され、論理ボリューム移動プラン情報 753 とファイルブロック移動プラン情報 754 が追加されたものとなっている。

## 【0179】

論理ボリューム移動プラン情報 753 には、移動順序 761、移動処理を行うホスト名 631、移動元ボリュームとその領域を指定する移動論理ボリューム名 773、移動論理ボリュームブロック番号 774、移動先の記憶装置 10 とその記憶領域を指定する移動先記憶装置名 764、移動先ボリューム名 765、移動先ブロック番号 766 の組が記憶される。ファイルブロック移動プラン情報 754 には、移動順序 761、移動処理を行うホスト名 651、移動元ファイルとその領域を指定するファイルシステム ID 535、移動データファイルパス名 775、移動ファイルブロック番号 776、移動先領域を指定する移動先記憶装置名 764、移動先ボリューム名 765、移動先ブロック番号 766 の組が記憶される。

## &lt;第三の実施の形態&gt;

本実施形態では、DBMS が実行される計算機とファイルを管理単位とする記憶装置がネットワークを用いて接続された計算機システムにおいて、データの記憶位置の管理を行う計算機が存在し、そこで計算機システム内のデータの記憶位置の管理を行う。計算機上で実行される OS 中のネットワークファイルシステムは、複数のファイルを仮想的な 1 つのファイルにまとめ、その構成を動的に変更する機能を有する。また、記憶装置においても、記憶装置内部でデータの記憶位置を動的に変更する機能を有する。

## 【0180】

データ記憶位置管理を実施する計算機は、DBMS に関する情報、データの記憶位置のマッピングに関する情報、記憶装置の構成情報を取得し、それらを用いて好適なデータ再配置案を作成する。ネットワークファイルシステム、記憶装置に対して作成したデータ配置を指示し、それらのデータ再配置機能を用いてそのデータ再配置案に従ったデータ配置を実現する。

## 【0181】

図 3 0 は、本発明の第三の実施の形態における計算機システムの構成図である。図示されたように、本発明の第三の実施の形態は本発明の第一の実施の形態と以下の点が異なる。

## 【 0 1 8 2 】

本実施の形態においては I / O パスインターフェイス 7 0、I / O バス 7 1、仮想ボリュームスイッチ 7 2 が存在せず、記憶制御装置 1 0 c と DB ホスト 8 0 c はネットワーク 7 9 を介してのみ接続される。記憶装置 1 0 はファイルを単位としたデータ記憶管理を行う記憶装置 1 0 c に変更される。そのため、物理記憶装置稼動情報 3 2、ボリューム物理記憶位置管理情報 3 6 がそれぞれ物理記憶装置稼動情報 3 2 c、ファイル記憶管理情報 3 6 c に変更される。

## 【 0 1 8 3 】

DB ホスト 8 0 c で実行される OS 1 0 0 ではボリュームマネージャ 1 0 2、ファイルシステム 1 0 4 が削除されその代わりにネットワークファイルシステム 1 0 4 c が追加され、OS 1 0 0 が保持するマッピング情報 1 0 6 がマッピング情報 1 0 6 c へ変更される。データ位置管理サーバ 8 2 内に記憶される記憶装置稼動情報 1 3 2、記憶装置構成情報 1 3 4、データ記憶位置管理情報 1 4 0 がそれぞれ記憶装置稼動情報 1 3 2 c、記憶装置構成情報 1 3 4 c、データ記憶位置管理情報 1 4 0 内のホストマッピング情報 6 5 0 の内容を若干変更したホストマッピング情報 6 5 0 c に変更される。

## 【 0 1 8 4 】

記憶装置 1 0 はファイルを管理単位とする記憶装置 1 0 c に変更される。DB ホスト 8 0 c からのアクセスも NFS 等のファイルをベースとしたプロトコルで実施される。記憶装置 1 0 におけるボリュームの役割は、記憶装置 1 0 c においてはファイルもしくはファイルを管理するファイルシステムとなり、そのファイルの記憶位置管理情報がファイル記憶管理情報 3 6 c である。1 つの記憶装置 1 0 c の中に複数のファイルシステムが存在しても構わない。

## 【 0 1 8 5 】

物理記憶装置 1 8 の稼動情報はボリュームを単位とした取得からファイルシステム又はファイルを単位とした取得に変更する。記憶装置 1 0 c 内にファイルシ

システムが存在する場合でもデータの移動機能を実現可能であり、データ移動指示方法は、前と同じく移動するファイルとそのデータ領域と、移動先領域を指定するものとする。本実施の形態においては、記憶装置 10c におけるデータ移動機能は必須であるとする。

## 【0186】

ネットワークファイルシステム 104c は、記憶装置 10c が提供するファイルをアクセスするための機能を提供する。更に、複数のファイルを仮想的な 1 つのファイルとして提供する。この機能を実現するためにネットワークファイルシステム 104c は管理情報をマッピング情報 106c 内に保持し、仮想ファイルアクセス時にこの管理情報を参照し、実際のアクセス先を求める処理を行う。更に、その構成を動的に変更する機能を有する。これら処理は、DB ホスト 80 上で管理コマンドを実行することにより実施される。

## 【0187】

図 31 は記憶装置 10c 内に保持される物理記憶装置稼動情報 32c を示す。図 2 の物理記憶装置稼動情報 32 からの変更点は、稼動情報取得単位がボリュームからファイルシステムに変更されたため、ボリューム名 501 の部分がファイルシステム名 1001 に変更されたことである。また、稼動情報取得単位をファイルとしてもよく、このときはボリューム名 501 の部分がファイルシステム名 1001 とファイルパス名 1002 に変更される。

図 32 は記憶装置 10c 内に保持されるファイル記憶管理情報 36c を示す。図 3 のボリューム物理記憶位置管理情報 36 からの変更点は、ボリューム物理記憶位置メイン情報 510、ボリュームデータ移動管理情報 511 からファイル物理記憶位置情報 510c、ファイルデータ移動管理情報 511c にそれぞれ変更される。上記の変更内容は、ボリュームの識別子がボリューム名 501 がファイルの識別子となるファイルシステム名 1001 とファイルパス名 1002 に、ボリューム内のデータ領域を示すボリューム論理ブロック番号 512 と移動論理ブロック番号 782 がそれぞれファイルブロック番号 1003 または移動ファイルブロック番号 1021 に変更されたことである。

## 【0188】

ここで、ファイルパス名 1 0 0 2 が “E m p t y” であるエントリ 1 0 1 5 は特殊なエントリであり、このエントリには記憶装置 1 0 c 内の物理記憶装置 1 8 の領域のうち、指定ファイルシステム内でファイルの記憶領域としてに割り当てられていない領域を示し、図 3 中のボリュームデータ移動管理情報 5 1 1 を用いるデータ移動方式で説明した処理手順を用い、この領域に対して移動するデータをコピーすることによりデータの物理記憶位置の動的変更機能を実現する。

## 【 0 1 8 9 】

図 3 3 は DB ホスト 8 0 c の OS 1 0 0 内に記憶されているマッピング情報 1 0 6 c を示す。マッピング情報 1 0 6 c 中には、ネットワークファイルシステムマウント情報 1 0 3 0 と仮想ファイル情報 1 0 4 0 と仮想ファイルデータ移動管理情報 1 0 5 0 が含まれる。

## 【 0 1 9 0 】

ネットワークファイルシステムマウント情報 1 0 3 0 は、記憶装置 1 0 c から提供され、DB ホスト 8 0 c においてマウントされているファイルシステムの情報で、ファイルシステムの提供元記憶装置とそのファイルシステムの識別子である記憶装置名 5 8 3 とファイルシステム名 1 0 0 1、そして、そのファイルシステムのマウントポイントの情報であるマウントポイント 1 0 3 1 の組を保持する。

## 【 0 1 9 1 】

仮想ファイル情報 1 0 4 0 は、ネットワークファイルシステム 1 0 4 c が提供する複数の記憶装置から提供されるファイルを仮想的な 1 つのファイルとして提供する機能の管理に用いる情報で、提供される仮想ファイルの識別子である仮想ファイルパス名 1 0 4 1 とそのデータ領域を示す仮想ファイルブロック番号 1 0 4 2 とそのデータ領域のデータを実際に保持するファイルの識別子である構成ファイルパス名 1 0 4 3 とその記憶領域を示す構成ファイルブロック番号 1 0 4 4 の組を含む。

## 【 0 1 9 2 】

仮想ファイルデータ移動管理情報は 1 0 5 0 は、ネットワークファイルシステム 1 0 4 c が提供する仮想ファイルの構成の変更処理の一部である、構成データ

の移動処理を行う際に利用する管理情報で、データ移動を行う移動元の仮想ファイルの識別子である移動仮想ファイルパス名 1 0 5 1 とデータ移動を行う領域を示す移動仮想ファイルブロック番号 1 0 5 2 とそのデータの移動先ファイルとその移動先領域を示す移動先構成ファイルパス名 1 0 5 3 と移動先ファイルブロック番号 1 0 5 4、そして、データ移動処理を行う際の管理情報である差分管理情報 7 8 5 とコピーポインタ 7 8 6 の組を含む。データの移動先に関しては、移動先ファイルの移動先指定領域に記憶領域実体が確保されている以外の制約は存在しない。本情報を利用し、図 3 中のボリュームデータ移動管理情報 5 1 1 を用いるデータ移動方式で説明した処理手順を用いることにより、データの移動機能を実現できる。

## 【 0 1 9 3 】

図 3 4 はデータ位置管理サーバ 8 2 c 上に記憶される記憶装置稼動情報 1 3 2 c を示す。図 8 の記憶装置稼動情報 1 3 2 からの変更点は、記憶装置稼動情報 1 3 2 ではボリューム単位で稼動情報を取得していたものを記憶装置稼動情報 1 3 2 c ではファイルシステム単位で稼動情報を取得していることである。そのため、ボリューム名 5 0 1 がファイルシステム名 1 0 0 1 に変更される。

## 【 0 1 9 4 】

図 3 5 はデータ位置管理サーバ 8 2 c 上に記憶される記憶装置構成情報 1 3 4 c を示す。図 9 の記憶装置構成情報 1 3 4 からの変更点は、記憶装置毎のボリューム物理記憶位置メイン情報 5 1 0 を記憶する記憶装置ボリューム物理記憶位置管理情報 6 0 4 が記憶装置毎のファイル物理記憶位置情報 5 1 0 c を記憶する記憶装置ファイル物理記憶位置情報 6 0 4 c に変更される。

## 【 0 1 9 5 】

図 3 6 はデータ位置管理サーバ 8 2 c 上に記憶されるホストマッピング情報 6 5 0 c を示す。図 1 2 のホストマッピング情報 6 5 0 からの変更点は、各ホストのマッピング情報 1 0 6 を保持するマッピング情報 6 5 2 が各ホストのマッピング情報 1 0 6 c を保持するマッピング情報 6 5 2 c になったことである。

## 【 0 1 9 6 】

データ再配置処理において、データ移動機能を保持する部分が仮想ボリューム

スイッチ 72 からネットワークファイルシステム 104 c に変更され、記憶装置 10 c がファイルを管理単位とするように変更されたことによる変更点は以下のようなものである。

【0197】

ネットワークファイルシステム 104 c においては、DB ホスト 80 c 上において管理コマンドを実施することによりデータの移動を実施する。そこで、データ位置管理主プログラム 130 がネットワーク 79 を通して、DB ホスト 80 c 上で実行されているデータ位置管理副プログラム 120 に、ネットワークファイルシステム 104 c においてデータの移動を実施する管理コマンドを実行する指示を出し、それに従ってデータ位置管理副プログラム 120 が管理コマンドを実施することによりデータの移動を実施する。このとき、現在空きの領域に対するデータの移動を行うため、データ移動先領域となるファイルやファイル内の領域が存在しない。このようなデータ移動指示を受け取った場合には、ネットワークファイルシステム 104 c は指定ファイルの新規ファイル作成や領域拡張を実施し、それが成功した後にデータ移動処理を開始し、データ移動中の領域不足の問題を回避する。

【0198】

また、ネットワークファイルシステム 104 c は通常のプロトコルを利用してファイル作成や領域拡張を実施する。そのため、そのデータ記憶先が必ずしも最適な場所に割り当てられるとは限らない。そこで、ネットワークファイルシステム 104 c によるデータ移動が完了後、今度は記憶装置 10 c に対して記憶装置内データ移動の指示を出し、作成したデータ再配置案に従ったデータ配置を実現する。このとき、記憶装置 10 c 内でデータ移動元と移動先が重なっている場合には、一旦データを移動先とは異なる空き領域にデータを移動し、その後に再度指定された移動先にデータを移動させる処理を記憶装置 10 c は実行する。

【0199】

データ再配置案作成処理においては、データ再配置案を作成する際に利用するワーク領域であるデータ再配置ワーク情報 670 がデータ再配置ワーク情報 670 c に、移動プラン情報 750 が移動プラン情報 750 c に変更される。

## 【0200】

図37にデータ再配置ワーク情報670cを示す。データ再配置ワーク情報670c中には、記憶装置ファイル物理記憶位置情報681cとデータ構造記憶装置内ファイル位置情報690cが含まれる。図14のデータ再配置ワーク情報670と比べて、仮想ボリューム物理記憶位置情報680は記憶装置ファイル物理記憶位置情報681cに変更され、データ構造仮想ボリューム内位置情報690はデータ構造仮想ボリューム内位置情報690cに変更される。

## 【0201】

記憶装置ファイル物理記憶位置情報681cは記憶装置10cが提供するファイルシステムとその内部に存在するファイル、その物理記憶装置18における記憶位置の一覧であり、記憶装置名583、ファイルシステム名1001、ファイルパス名1002、ファイルブロック番号1003、物理記憶装置名502、物理ブロック番号514の組を保持する。このデータは、記憶装置構成情報134cを参照して初期化する。ファイルパス名1002が“Empty”であるエントリ1071は記憶装置10cのファイルシステム名1001の領域のうち、ファイルの記憶に利用されていない領域を示し、ここに対してデータの移動が可能である。

## 【0202】

データ構造仮想ボリューム内位置情報690cはDBMS110が保持しているデータ構造が記憶装置10cが提供するファイルのどこに記憶されているかを示した情報であり、ホスト名651、DBMS名631、データ構造名561、データファイルパス名562、ファイルブロック番号563、記憶装置名583、ファイルシステム名1001、ファイルパス名1002、ファイルブロック番号1003の組を保持する。この情報は、DBMSスキーマ情報136中のDBMSデータ記憶位置情報622とDBMSホスト情報626とホストマッピング情報650cを参照し、ファイルパスに関して対応する部分をまとめることにより初期データを作成する。

## 【0203】

図38に移動プラン情報750cを示す。移動プラン情報750cは、ネット

ワークファイルシステム 1 0 4 c に対するデータ移動指示を記憶する仮想ファイルブロック移動プラン情報 7 5 5 と記憶装置 1 0 c に対するデータ移動指示を記憶する物理記憶位置移動プラン情報 7 5 2 c を含む。これら情報に関しては、何もデータを持たないように初期化する。図 1 5 の移動プラン情報 7 5 0 と比較して、移動プラン情報 7 5 0 c は仮想ボリューム移動プラン情報 7 5 2 が削除され、仮想ファイル移動プラン情報 7 5 5 が追加され、物理記憶位置移動プラン情報 7 5 2 が物理記憶位置移動プラン情報 7 5 2 c に変更されたものとなる。

#### 【 0 2 0 4 】

仮想ファイル移動プラン情報 7 5 5 には、移動順序 7 6 1、移動処理を行うホスト名 6 5 1、移動元の仮想ファイルとその領域を指定する移動仮想ファイルパス名 1 0 5 1、移動仮想ファイルブロック番号 1 0 5 2、移動先の構成ファイルとその領域を指定する移動先構成ファイルパス名 1 0 5 3、移動先ファイルブロック番号 1 0 5 4 の組が記憶される。物理記憶位置移動プラン情報 7 5 2 c には、移動順序 7 6 1、移動処理を行う移動記憶装置名 7 6 7、移動元のファイルとその領域を指定する移動ファイルシステム名 1 1 0 1、移動ファイルパス名 1 1 0 2、移動ファイルブロック番号 1 1 0 3、移動先の物理記憶装置 1 8 とその領域を指定する移動先物理記憶装置名 7 7 1、移動先物理ブロック番号 7 7 2 の組が記憶される。

#### 【 0 2 0 5 】

本実施例においては、ネットワークファイルシステム 1 0 4 c が複数のファイルから 1 つの仮想ファイルを構成する機能を有しているが、ここで、単純に 1 つの仮想ファイルは 1 つのファイルから構成されても構わない。このとき、ネットワークファイルシステム 1 0 4 c は動的なデータ移動機能のみを提供する。さらに、DBMS 1 1 0 が処理を中断しても構わない場合には、DBMS 1 1 0 の処理の中断後、ファイルのコピーを実施し、アクセス先がコピー先になるようにシンボリックリンクを張りなおした後に DBMS 1 1 0 が処理を再開することによるデータ再配置も可能である。さらに、ネットワークファイルシステム 1 0 4 c におけるデータ移動を行わず、記憶装置 1 0 c のみでデータ移動を実施することも可能である。



## 【 0 2 0 6 】

## 【発明の効果】

本発明により以下のことが可能となる。第一に、DBMSが管理するデータの特性を考慮して記憶装置におけるデータ記憶位置を決定することにより、DBMSに対してより好ましいアクセス性能特性を持つ記憶装置を保持する計算機システムが実現される。これにより、その計算機システムで稼動しているDBMSの性能を向上させることができる。特に、複数の記憶装置を利用するDBシステムにおいて、各記憶装置間へアクセス要求が適切に分散化されるようにし、DBMSの処理性能を向上させる。

## 【 0 2 0 7 】

第二に、DBMSが稼動している計算機システムにおいて、DBMSの特性を考慮した記憶装置の良好なアクセス性能特性達成を目的としたデータ記憶位置再配置処理を実現するため、計算機システムの性能に関する管理コストを削減することができる。特に、本発明を用いることにより、データ記憶位置の再配置案を自動で作成することが可能であり、管理コストの削減に大きく寄与する。更に、多数のDBMSが稼動し、多数の記憶装置が存在するシステムにおいても利用可能で集中的な管理を可能とするため、そのようなシステムにおける性能に関する管理コストを削減することができる。

## 【図面の簡単な説明】

## 【図 1】

第一の実施の形態における計算機システムの構成を示す図である。

## 【図 2】

記憶装置 10 内に保持されている物理記憶装置稼動情報 32 を示す図である。

## 【図 3】

記憶装置 10 内に保持されているボリューム物理記憶位置管理情報 36 を示す図である。

## 【図 4】

第一の実施の形態におけるDBホスト 80 のOS 100 内に記憶されているマッピング情報 106 を示す図である。

【図 5】

DBMS 1 1 0 内に記憶されているその内部で定義・管理しているデータその他の管理情報であるスキーマ情報 1 1 4 を示す図である。

【図 6】

DB ホスト 8 0 のメモリ 8 8 上に記憶されている実行履歴情報 1 2 2 を示す図である。

【図 7】

仮想ボリュームスイッチ 7 2 が保持する仮想ボリューム情報 7 3 を示す図である。

【図 8】

データ位置管理サーバ 8 2 上に記憶される記憶装置稼動情報 1 3 2 を示す図である。

【図 9】

データ位置管理サーバ 8 2 上に記憶される記憶装置構成情報 1 3 4 を示す図である。

【図 1 0】

データ位置管理サーバ 8 2 上に記憶される DBMS スキーマ情報 1 3 6 を示す図である。

【図 1 1】

データ位置管理サーバ 8 2 上に記憶される DBMS 実行履歴情報 1 3 8 を示す図である。

【図 1 2】

データ位置管理サーバ 8 2 上に記憶されるデータ記憶位置管理情報 1 4 0 を示す図である。

【図 1 3】

データ位置管理主プログラム 1 3 0 によるデータ再配置処理の処理フローを示す図である。

【図 1 4】

データ配置解析・再配置案作成処理で利用するデータ再配置ワーク情報 6 7 0

を示す図である。

【図 1 5】

データ配置解析・再配置案作成処理により作成されるデータ移動案を格納する移動プラン情報 7 5 0 を示す図である。

【図 1 6】

記憶装置稼動情報 1 3 2 を基にした同時アクセス実行データ構造を分離するためのデータ再配置案作成処理の処理フローを示す図である。

【図 1 7】

DBMS 実行履歴情報 1 3 8 を基にした同時アクセス実行データ構造を分離するためのデータ再配置案作成処理で利用するクエリ実行時同時アクセスデータカウント情報 7 0 0 を示す図である。

【図 1 8】

DBMS 実行履歴情報 1 3 8 を基にした同時アクセス実行データ構造を分離するためのデータ再配置案作成処理処理フローを示す図である。

【図 1 9】

指定されたデータ構造とそのデータ構造と同時にアクセスされる可能性が高いデータ構造の組を分離するデータ再配置案を作成する処理のフローを示す図である。

【図 2 0】

データ構造の定義を基にした同時アクセス実行データ構造を分離するためのデータ再配置案作成処理の処理フローを示す図である。

【図 2 1】

特定の表や索引の同一データ構造に対するアクセス並列度を考慮したデータ再配置案作成処理の処理フローを示す図である。

【図 2 2】

特定の表データに対するシーケンシャルアクセス時のディスクネックを解消するデータ再配置案作成処理の処理フローを示す図である。

【図 2 3】

特定のデータ構造に対する記憶装置 1 0 におけるキャッシュ効果を考慮したデ

ータ再配置案作成処理で利用するキャッシュ効果判定処理の処理フローを示す図である。

【図 2 4】

特定のデータ構造に対する記憶装置 1 0 におけるキャッシュ効果を考慮したデータ再配置案作成処理の処理フローを示す図である。

【図 2 5】

第二の実施の形態における計算機システムの構成を示す図である。

【図 2 6】

DBホスト 8 0 b の OS 1 0 0 内に記憶されるマッピング情報 1 0 6 b を示す

【図 2 7】

データ位置管理サーバ 8 2 上に記憶されるホストマッピング情報 6 5 0 b を示す図である。

【図 2 8】

データ配置解析・再配置案作成処理で利用するデータ再配置ワーク情報 6 7 0 b を示す図である。

【図 2 9】

データ配置解析・再配置案作成処理により作成されるデータ移動案を格納する移動プラン情報 7 5 0 b を示す図である。

【図 3 0】

第三の実施の形態における計算機システムの構成を示す図である。

【図 3 1】

記憶装置 1 0 c 内に保持される物理記憶装置稼動情報 3 2 c を示す図である。

【図 3 2】

記憶装置 1 0 c 内に保持されるファイル記憶管理情報 3 6 c を示す図である。

【図 3 3】

DBホスト 8 0 c の OS 1 0 0 内に記憶されているマッピング情報 1 0 6 c を示す図である。

【図 3 4】

データ位置管理サーバ 82 c 上に記憶される記憶装置稼動情報 132 c を示す図である。

【図 35】

データ位置管理サーバ 82 c 上に記憶される記憶装置構成情報 134 c を示す図である。

【図 36】

データ位置管理サーバ 82 c 上に記憶されるホストマッピング情報 650 c を示す図である。

【図 37】

データ配置解析・再配置案作成処理で利用するデータ再配置ワーク情報 670 c を示す図である。

【図 38】

データ配置解析・再配置案作成処理で利用する移動プラン情報 750 c を示す図である。

【符号の説明】

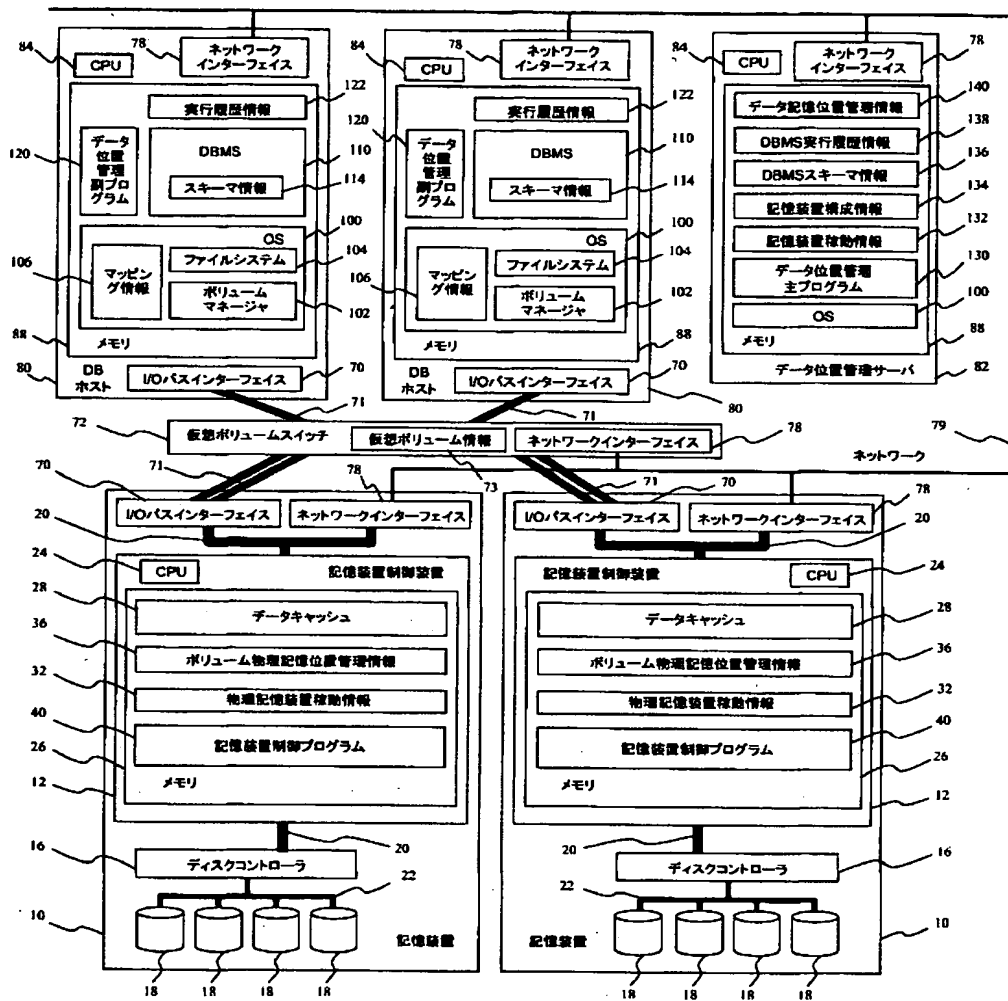
10, 10 c	記憶装置
18	物理記憶装置
28	データキャッシュ
32, 32 c	物理記憶装置稼動情報
36	ボリューム物理記憶位置管理情報
36 c	ファイル記憶管理情報
70	I/Oパスインターフェイス
71	I/Oパス
72	仮想ボリュームスイッチ
72 b	I/Oパススイッチ
73	仮想ボリューム情報
78	ネットワークインターフェイス
79	ネットワーク
80, 80 b, 80 c	DBホスト

82	データ位置管理サーバ
100	OS (オペレーティングシステム)
102, 102b	ボリュームマネージャ
104, 104b	ファイルシステム
104c	ネットワークファイルシステム
106, 106b, 106c	マッピング情報
110	DBMS (データベース管理システム)
114	スキーマ情報
120	データ位置管理副プログラム
122	実行履歴情報
130	データ位置管理主プログラム
132, 132c	記憶装置稼動情報
134, 134c	記憶装置構成情報
136	DBMS スキーマ情報
138	DBMS 実行履歴情報
140	データ記憶位置管理情報
650, 650b, 650c	ホストマッピング情報

【書類名】 図面

【図1】

図1



【図 2】

図2

501 ボリューム名	502 物理記憶装置名	503 累積移動時間
Vol0-0	Pdisk0-0	23917390
	Pdisk0-1	38902849
Vol0-1	Pdisk0-0	8290181
	Pdisk0-1	13908202

物理記憶装置移動情報

32

【図 3】

図3

501 ボリューム名	512 ボリューム論理 ブロック番号	502 物理記憶 装置名	514 物理ブロック番号
Vol 0-0	0 - 10239	Pdisk 0-0	0 - 10239
	10240 - 20479	Pdisk 0-1	0 - 10239
Vol 0-1	0 - 10239	Pdisk 0-0	10240 - 20479
	10240 - 20479	Pdisk 0-1	10240 - 20479
⋮	⋮	⋮	⋮
Empty	-	Pdisk 0-0	40960 - 122880
	-	Pdisk 0-1	40960 - 122880

ボリューム物理記憶位置メイン情報

510

515

501 ボリューム 名	782 移動論理 ブロック番号	783 移動先物理 記憶装置名	784 移動先物理 ブロック番号	785 差分管理情報	786 コピー ポインタ
Vol 0-1	8192 - 10239	Pdisk 0-1	38912 - 40959	0 1 ... 0	9840
Vol 0-2	0 - 8191	Pdisk 0-1	30720 - 38911	0 0 ... 1	1792
⋮	⋮	⋮	⋮	⋮	⋮

ボリュームデータ移動管理情報

511

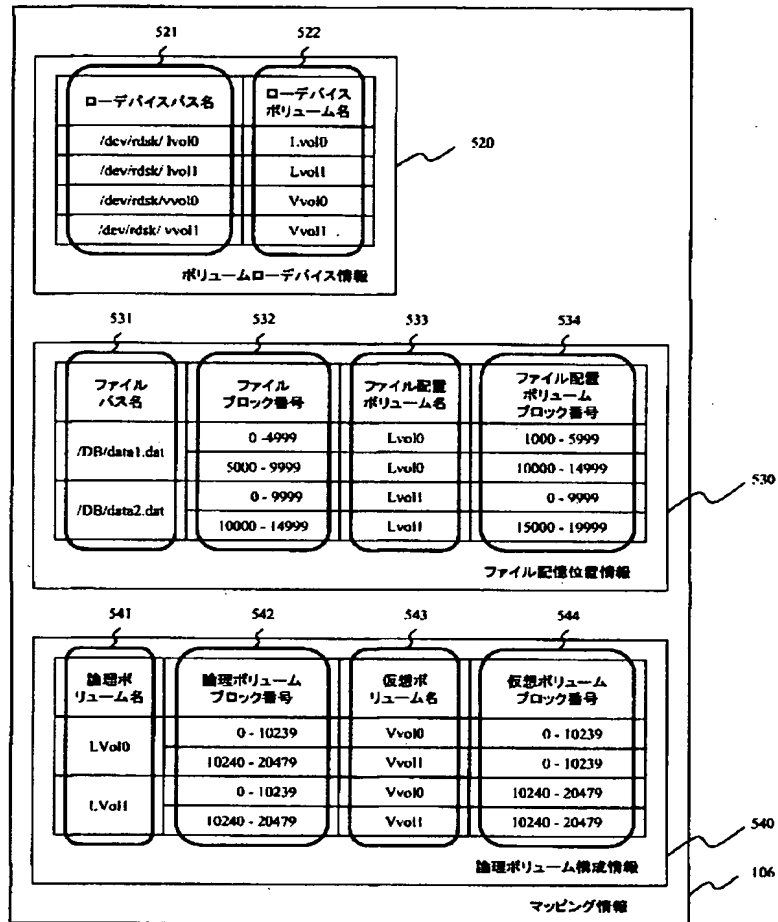
36

ボリューム物理記憶位置管理情報



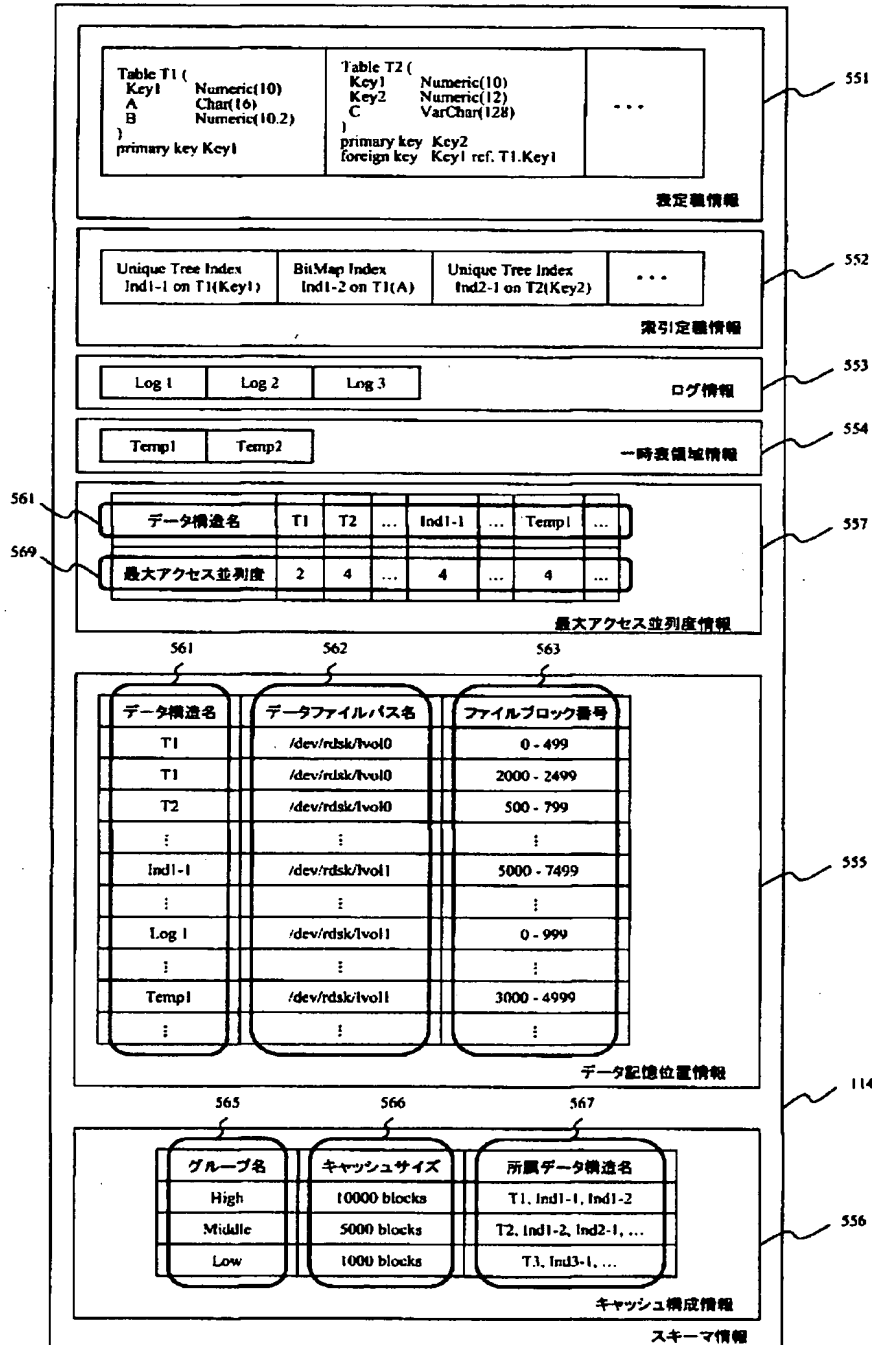
【図4】

図4



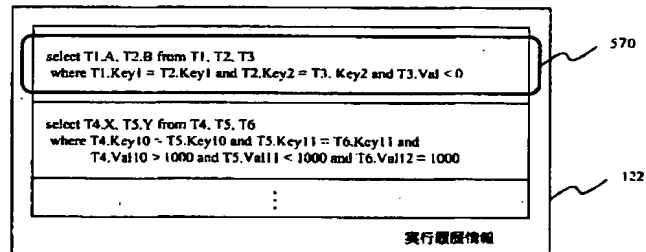
【図 5】

図5



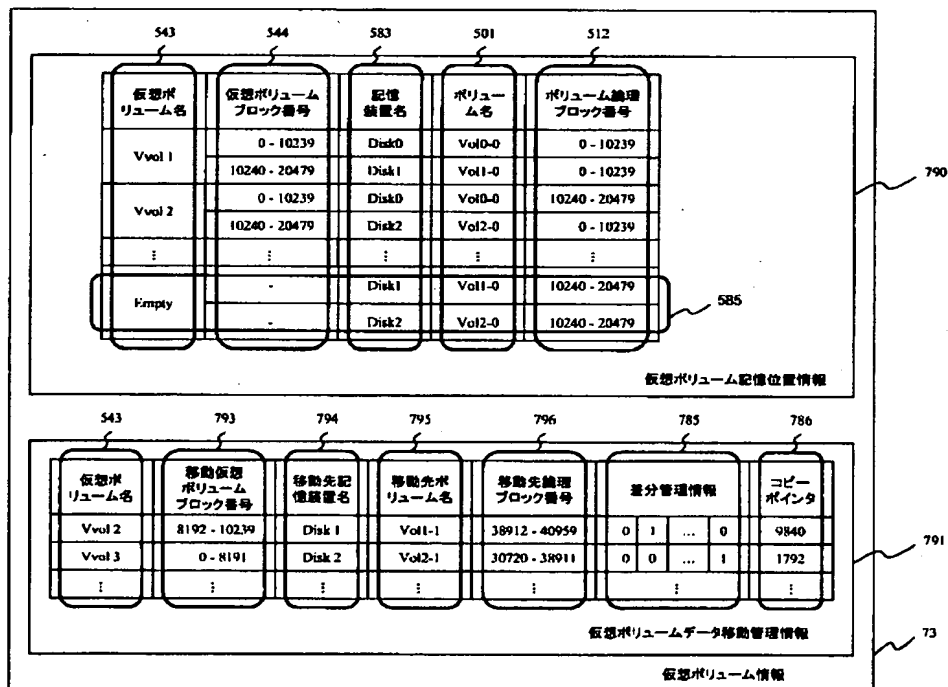
【図 6】

図6



【図 7】

図7



【図 8】

図8

記憶装置名		Disk0	Disk0	Disk0	...	Disk2	...
ボリューム名		Vol0-0	Vol0-0	Vol0-1	...	Vol2-0	...
物理記憶装置名		Pdisk 0-0	Pdisk 0-1	Pdisk 0-0	...	Pdisk 2-0	...
旧累積稼働時間		23917390	38902849	8012891	...	128928479	...
稼働率	2000/4/1 12:00 ~ 2000/4/1 12:15	20%	12%	4%	...	50%	...
	2000/4/1 12:15 ~ 2000/4/1 12:30	15%	10%	7%	...	40%	...
	2000/4/1 12:30 ~ 2000/4/1 12:45	16%	9%	5%	...	43%	...
	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	⋮	⋮	⋮	⋮	⋮	⋮	⋮
前回累積稼働時間取得時刻: 2001/4/12 18:15		記憶装置稼働情報					

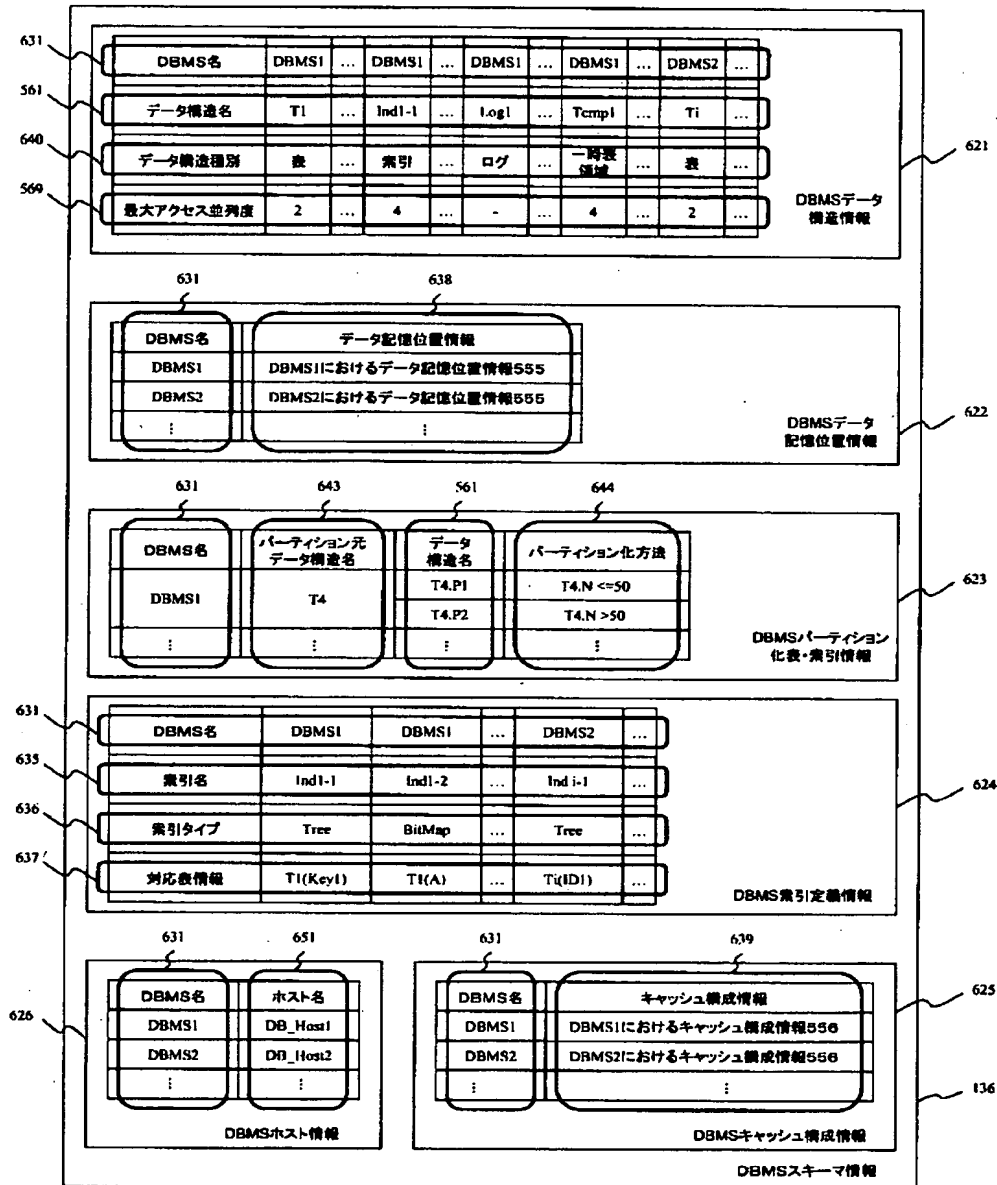
【図 9】

図9

583	601	602	603
記憶装置名	移動稼働情報	データキャッシュ容量	記憶装置ボリューム物理記憶位置管理情報
Disk0	有	4096MB	Disk0におけるボリューム物理記憶位置メイン情報510
Disk1	無	-	既知ルールに従った記憶位置情報
⋮	⋮	⋮	⋮
記憶装置構成情報			

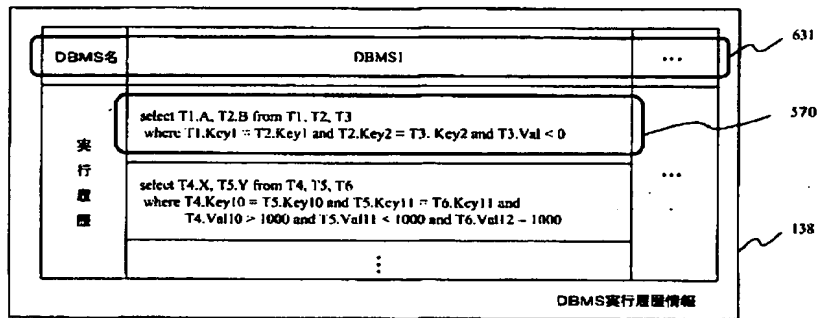
【図10】

図10



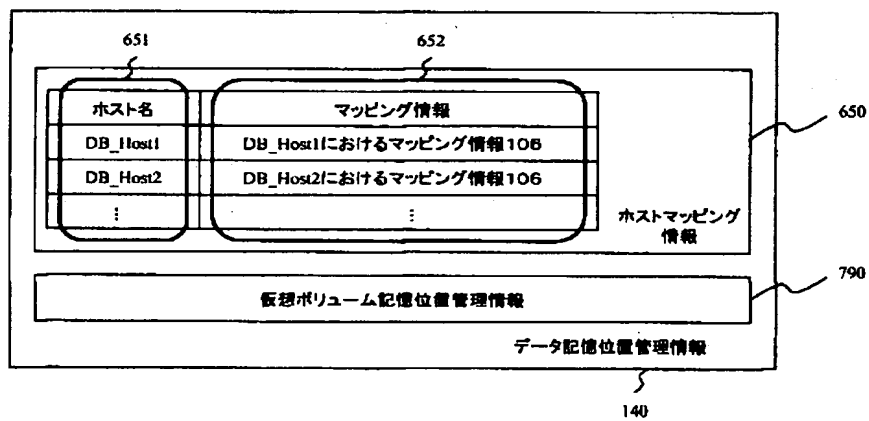
【図 1 1】

図11



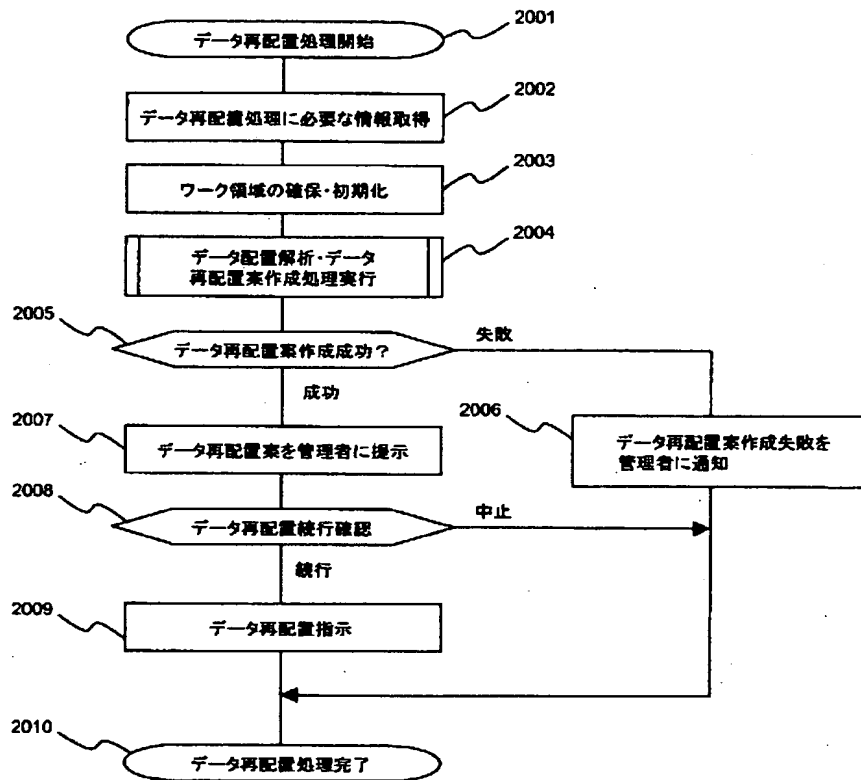
【図 1 2】

図12



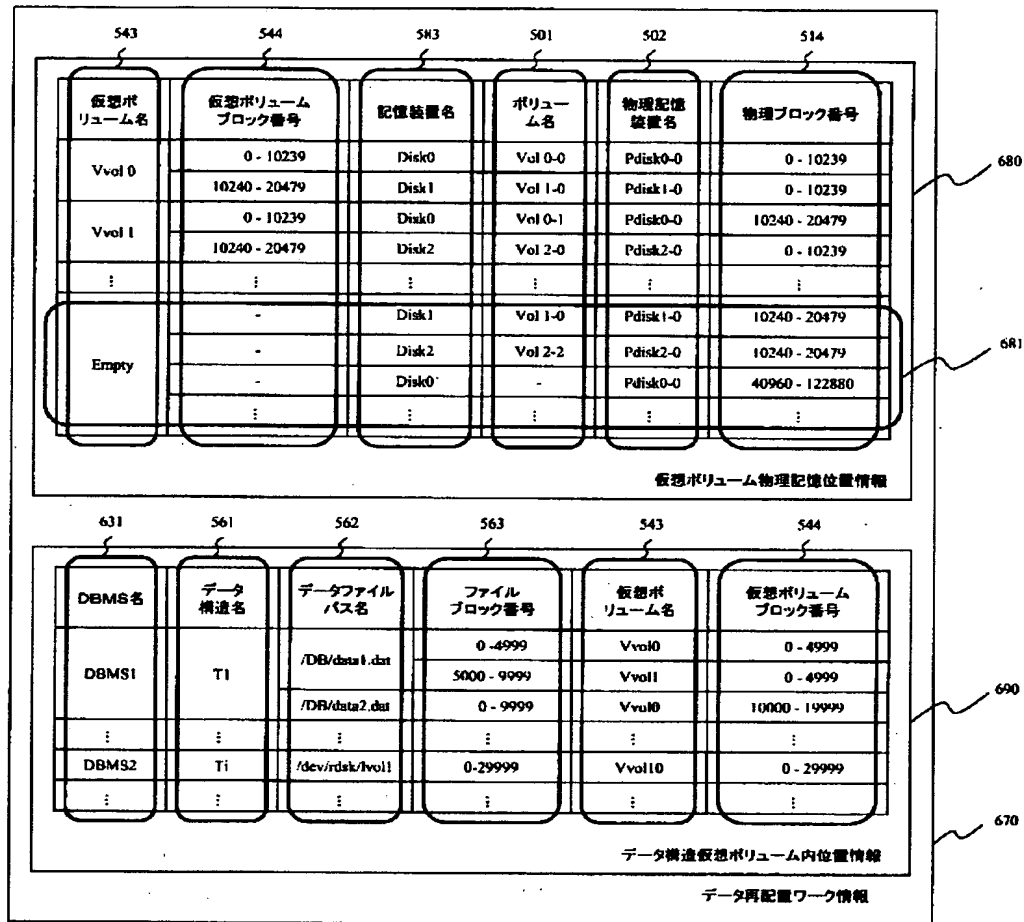
【図13】

図13



【図14】

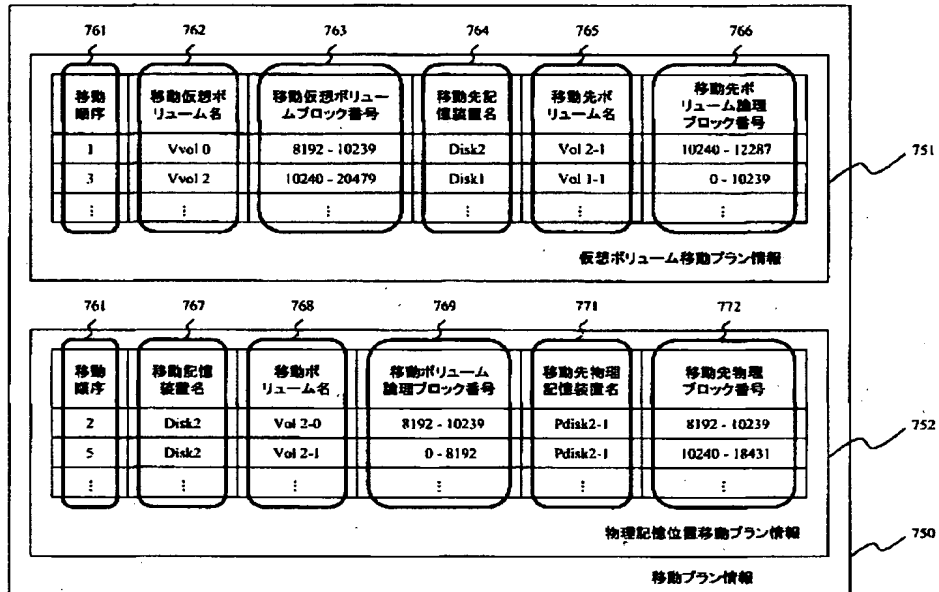
図14



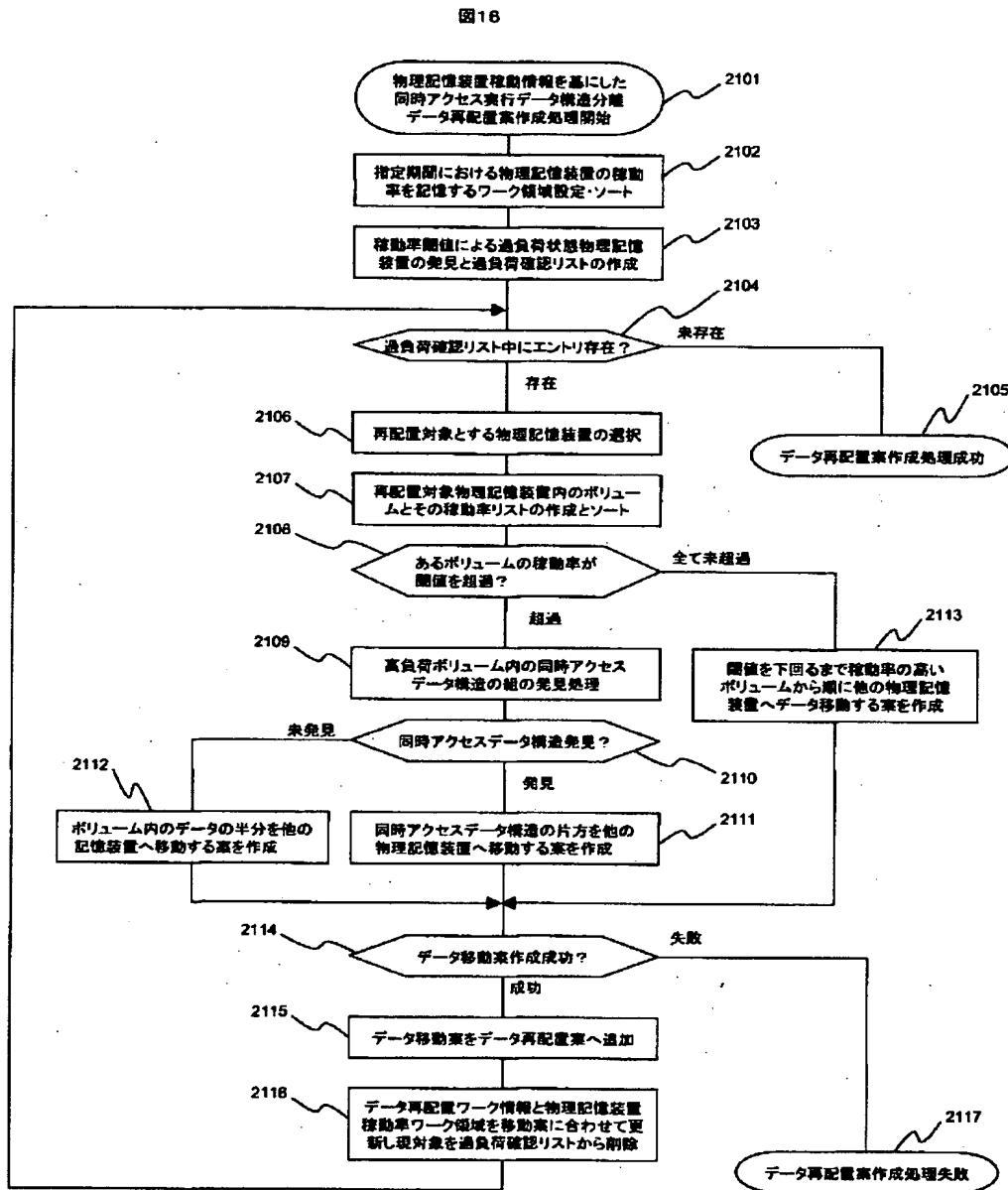


【図15】

図15



【図 16】



【図 17】

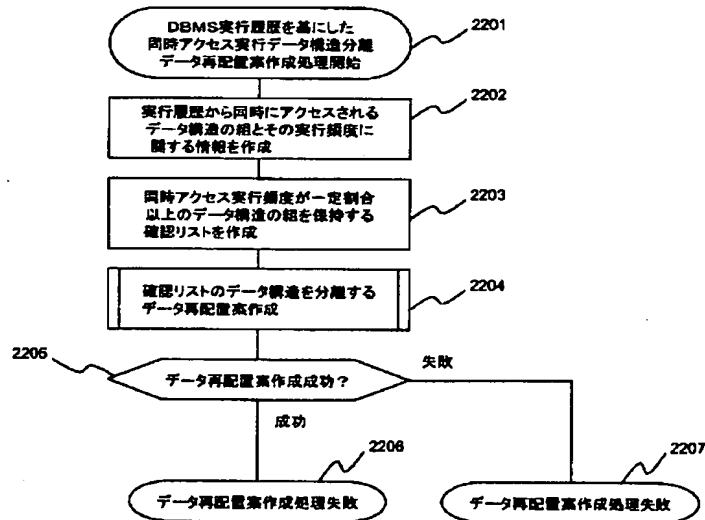
図 17

631	701	702	703	
DBMS名	データ構造名A	データ構造名B	カウント値	
DBMS1	T1	Ind1-1	2789	
⋮	⋮	⋮	⋮	
DBMS2	Ti	Ind i-1	829	
⋮	⋮	⋮	⋮	

クエリ実行同時アクセスデータ構造カウント情報

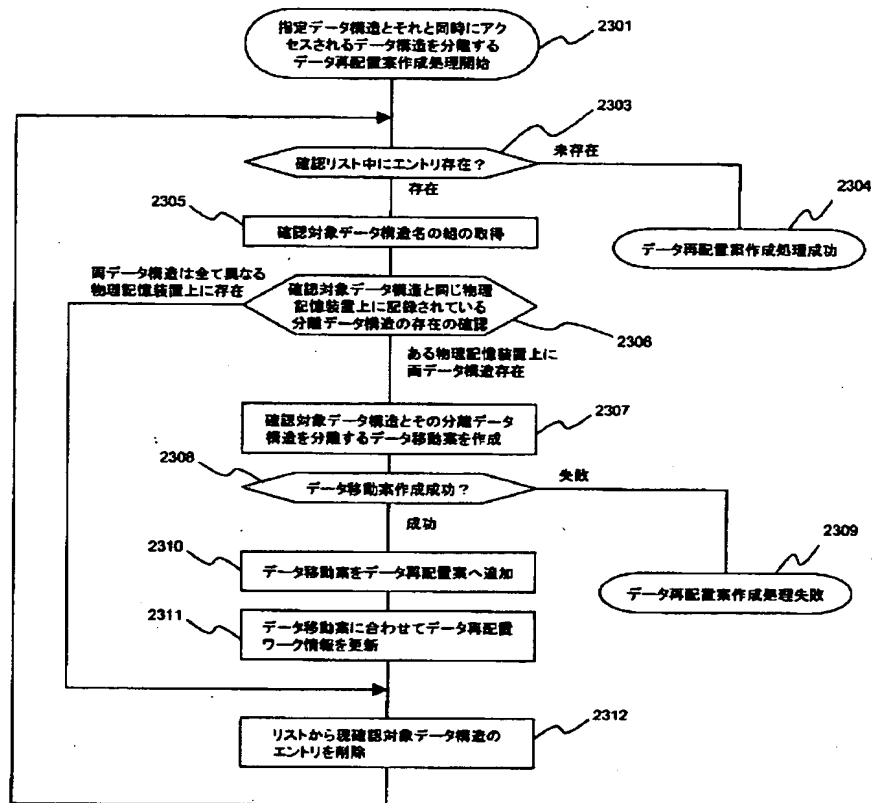
【図 18】

図 18

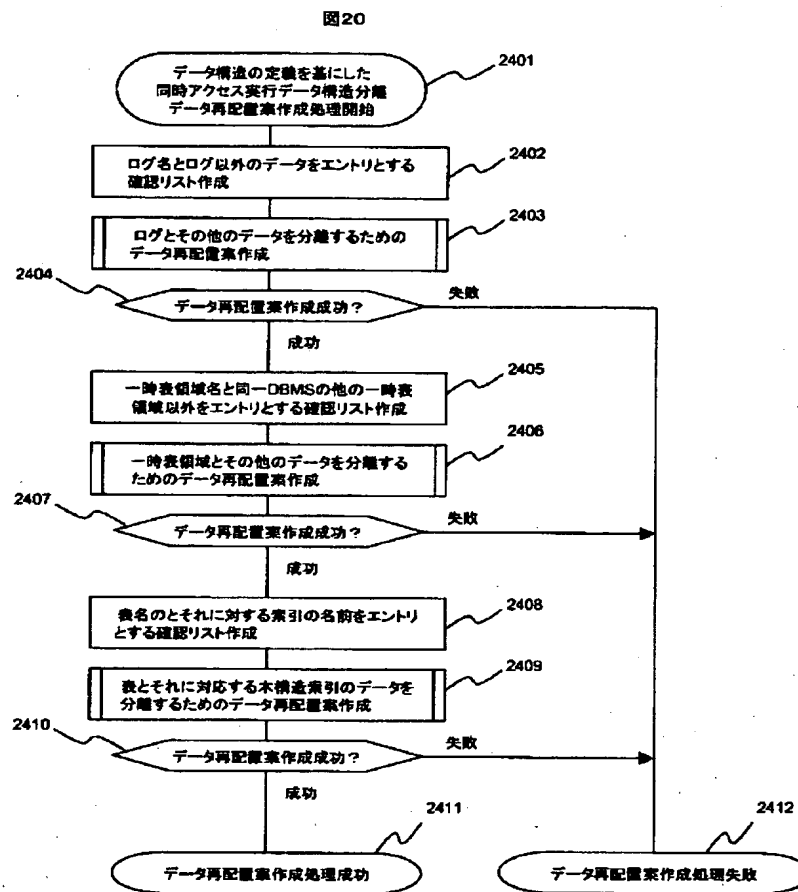


【図19】

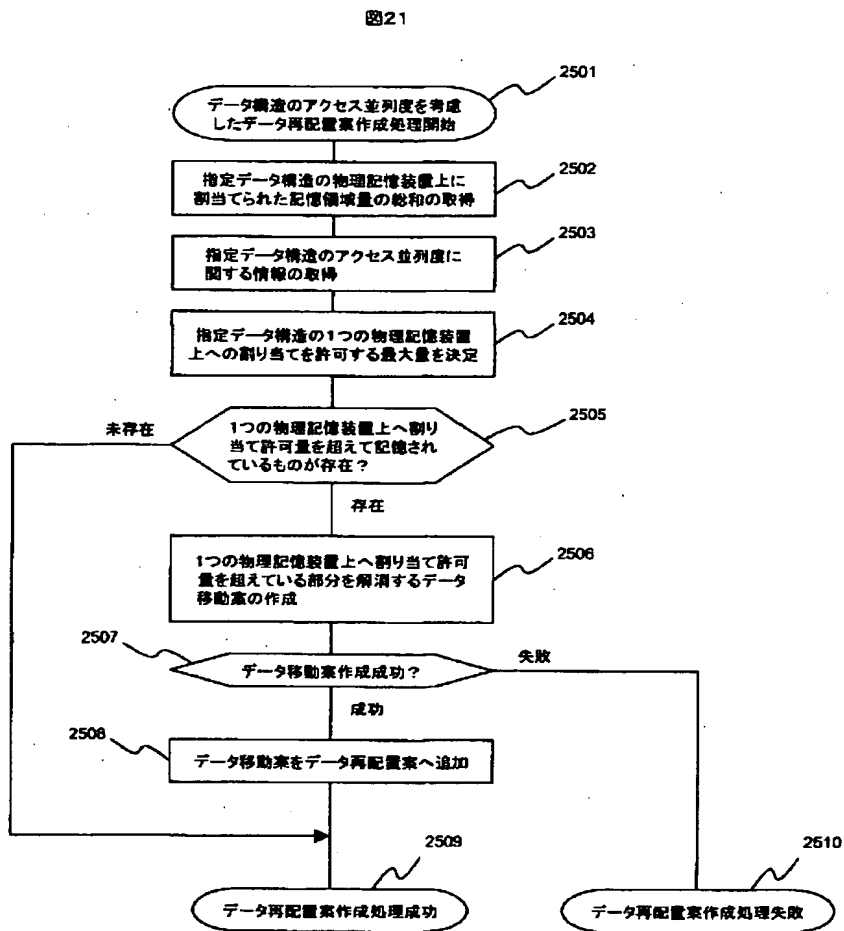
図19



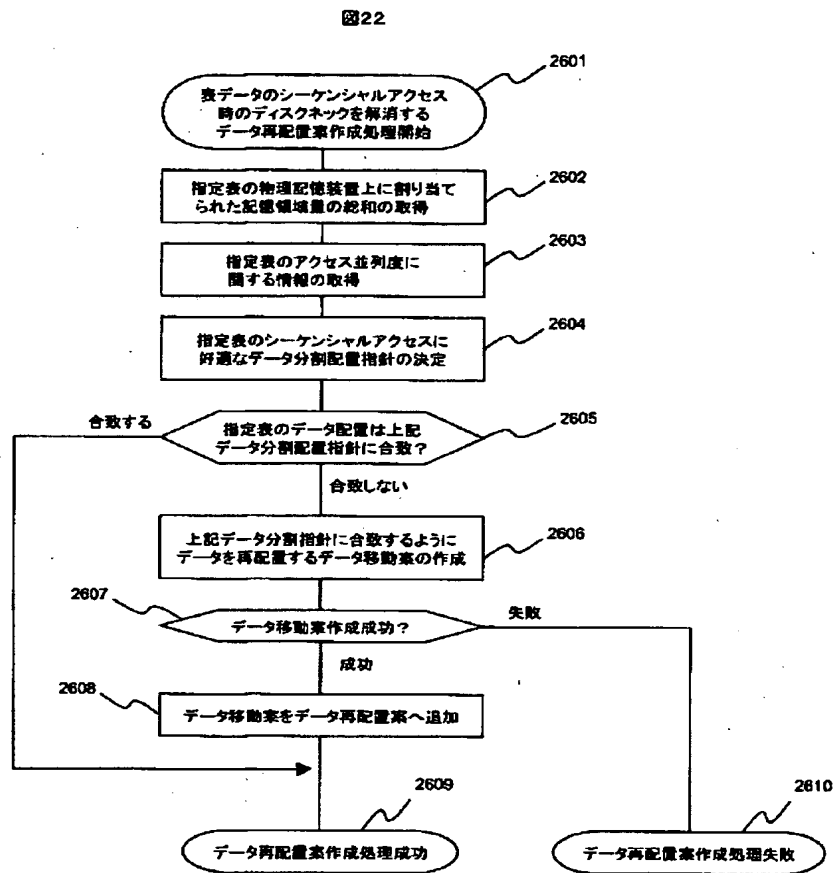
【図20】



【図 21】

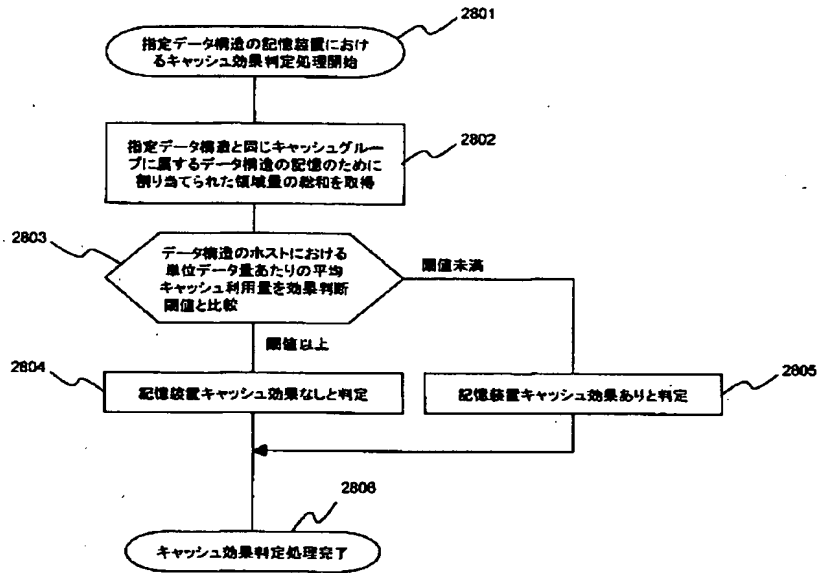


【図 22】



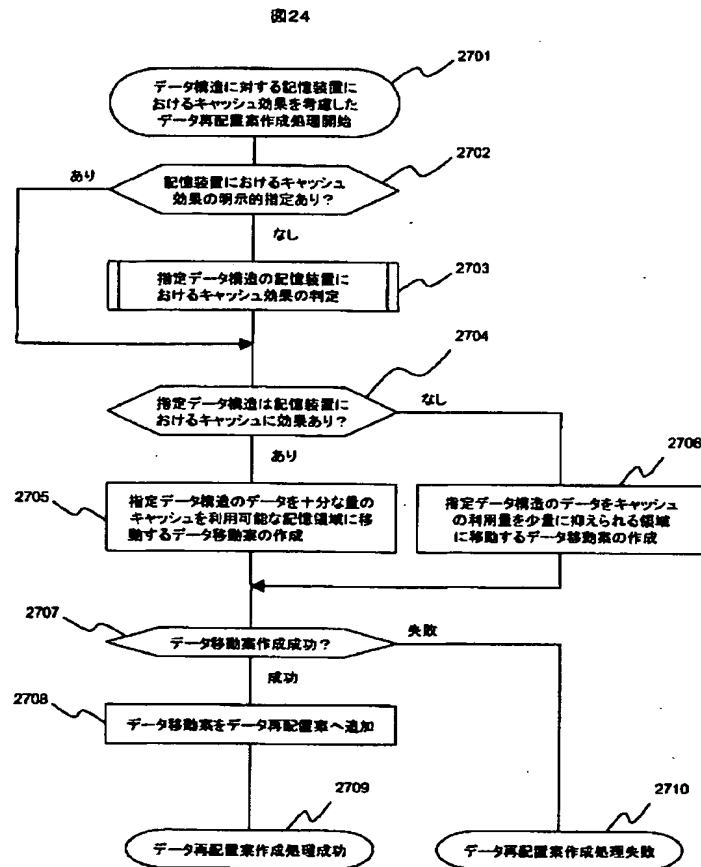
【図 23】

図23



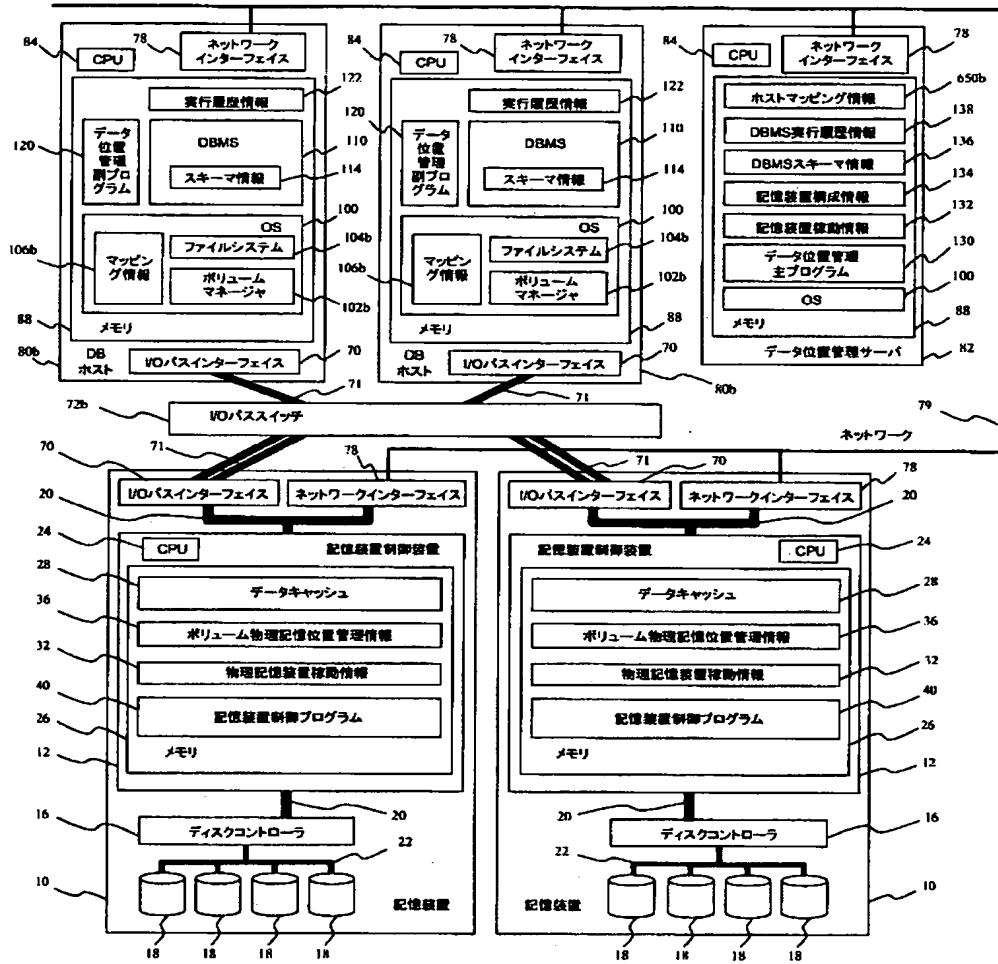


【図24】



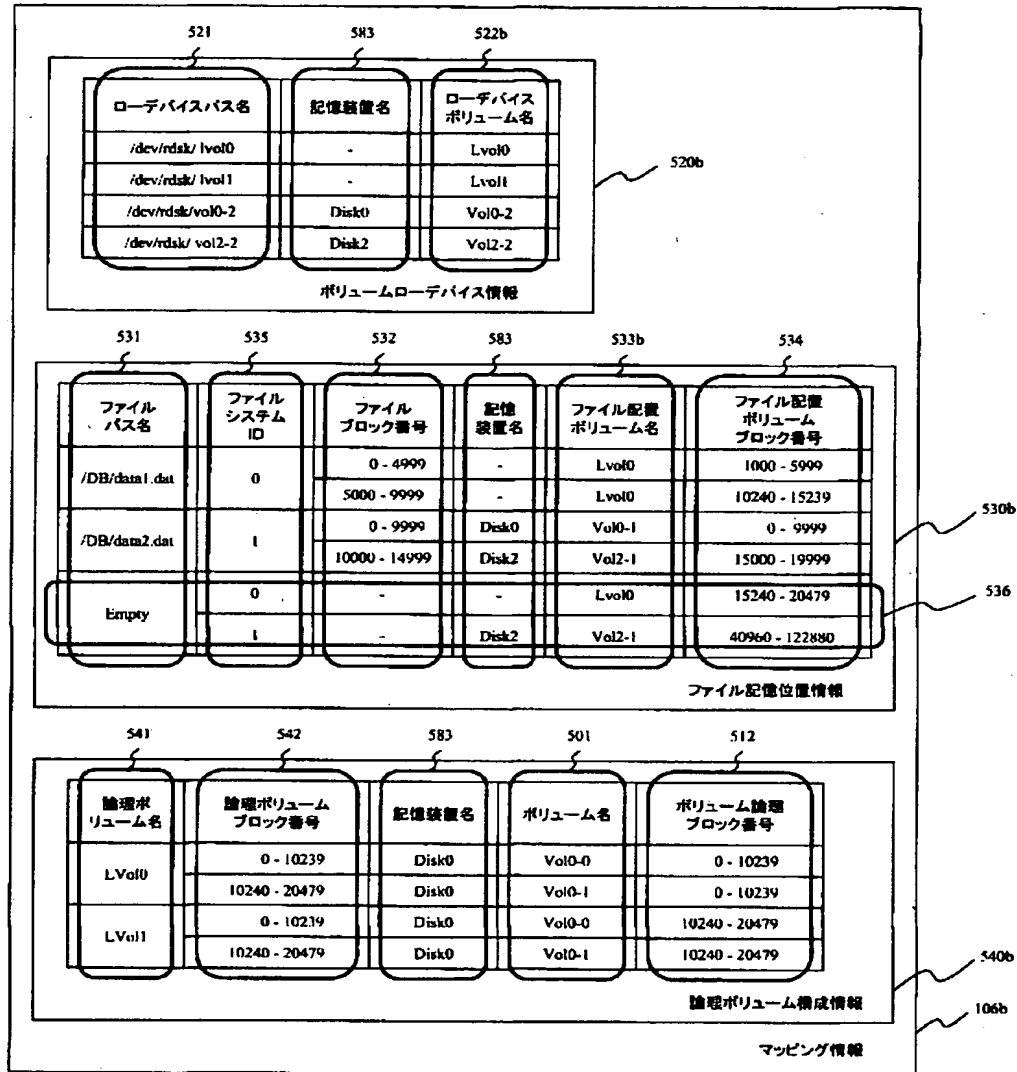
【図 25】

図26



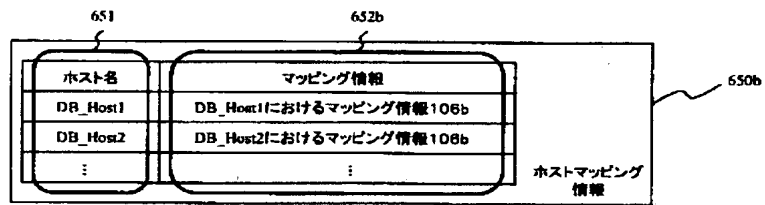
【図26】

図26



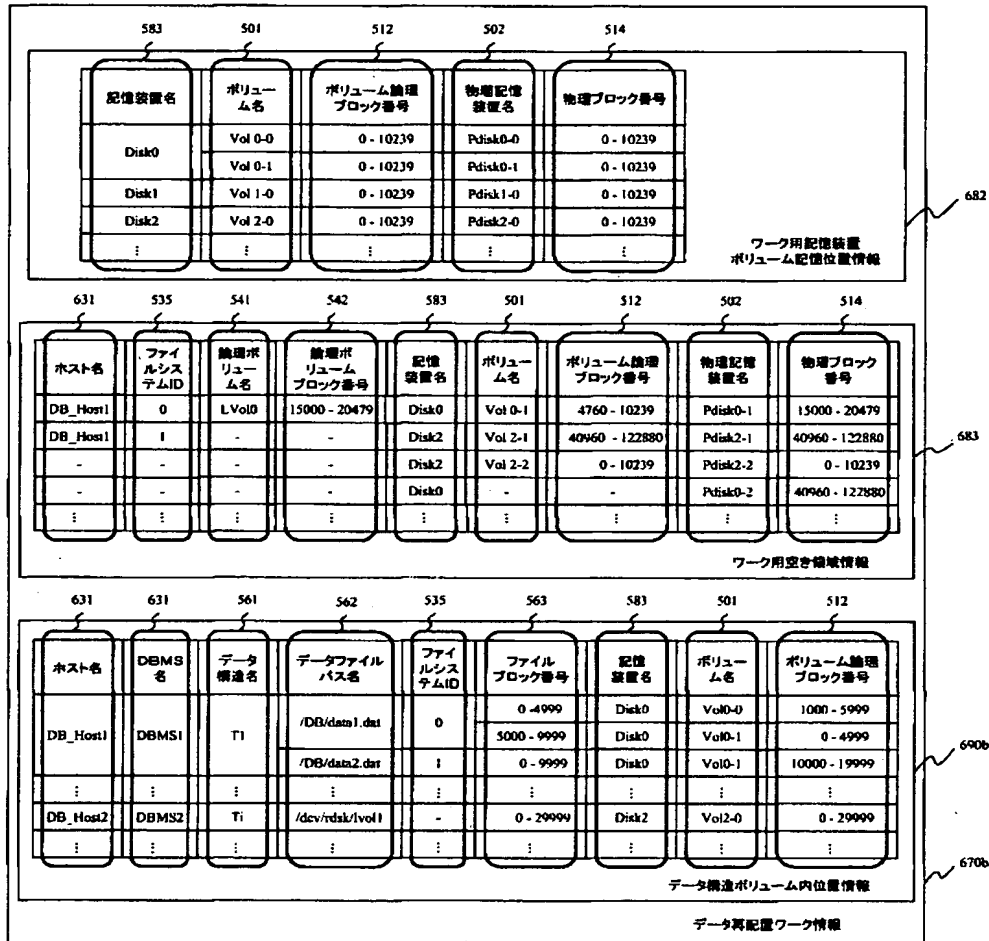
【図 2 7】

図27



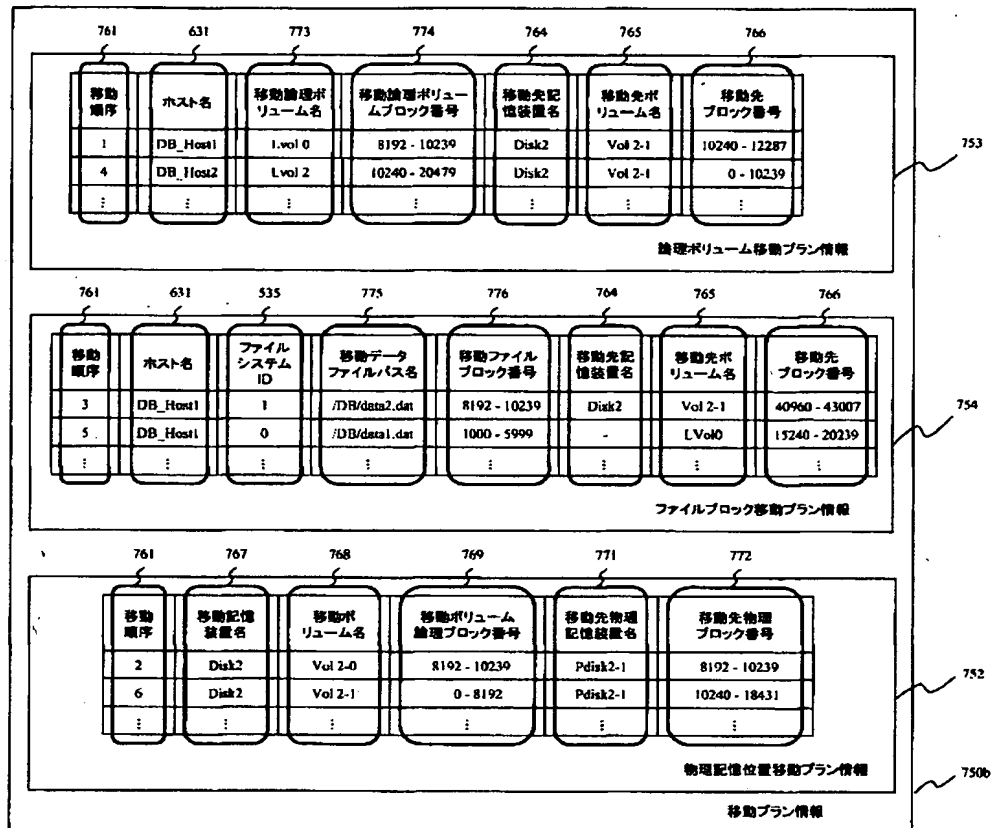
【図 28】

図28



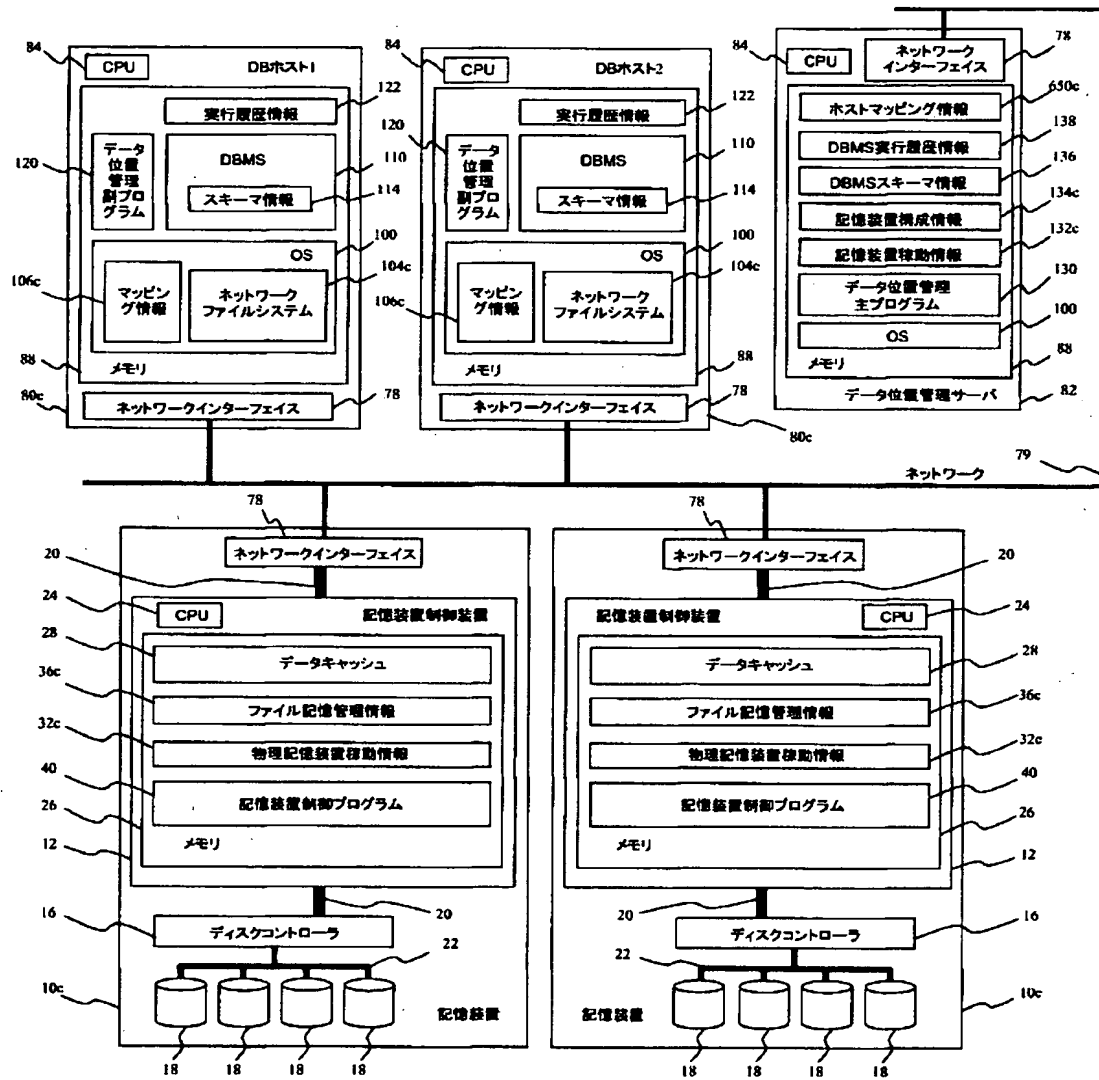
【図 29】

図29



【図30】

図30



【図 3 1】

図31

1001 ファイル システム名	502 物理記憶装置名	503 累積稼働時間
FS0-0	Pdisk0-0	23917390
	Pdisk0-1	38902849
FS0-1	Pdisk0-0	8290181
	Pdisk0-1	13908202

32c

物理記憶装置稼働情報

【図 3 2】

図32

1001 ファイル システム名	1002 ファイル パス名	1003 ファイル ブロック番号	502 物理記憶 装置名	514 物理ブロック番号
FS 0-0	/control.dat	0 - 1023	Pdisk 0-0	4096 - 5119
	/VF/VF1.dat	0 - 2047	Pdisk 0-1	1024 - 3071
	/VF/VF1.dat	2048 - 4095	Pdisk 0-0	8192 - 10239
	Empty	-	Pdisk 0-0	40960 - 122880
FS 0-1	/VF/VF3.dat	0 - 10239	Pdisk 0-1	10240 - 20479
⋮	⋮	⋮	⋮	⋮

510c

1015

ファイル物理記憶位置情報

1001 ファイル システム名	1002 ファイル パス名	1021 移動ファイル ブロック番号	783 移動先物理 記憶装置名	784 移動先物理 ブロック番号	785 差分管理情報	786 コピー ポイント
FS 0-1	/data1.dat	8192 - 10239	Pdisk 0-1	38912 - 40959	0 1 ... 0	9840
FS 0-2	/dataA.dat	0 - 8191	Pdisk 0-1	30720 - 38911	0 0 ... 1	1792
⋮	⋮	⋮	⋮	⋮	⋮	⋮

511c

36c

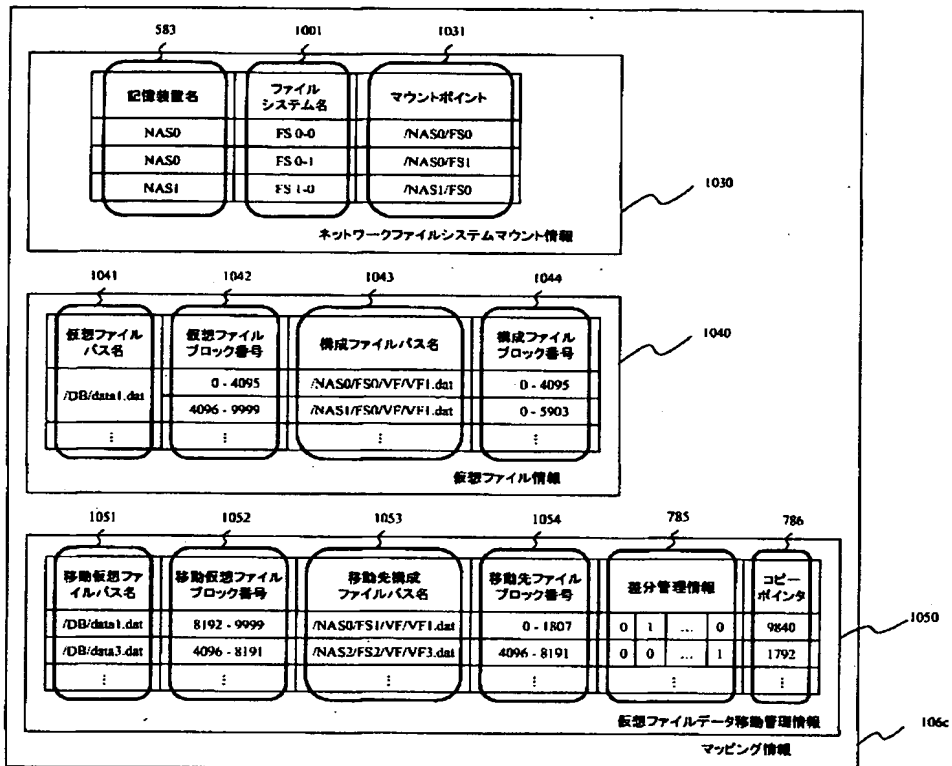
ファイルデータ移動管理情報

ファイル記憶管理情報



【図 33】

図33



【図 34】

図34

記憶装置名		Disk0	Disk0	Disk0	...	Disk2	...
ファイルシステム名		Vol0-0	Vol0-0	Vol0-1	...	Vol2-0	...
物理記憶装置名		Pdisk 0-0	Pdisk 0-1	Pdisk 0-0	...	Pdisk 2-0	...
旧累積移動時間		23917390	38902849	8012891	...	128928479	...
操 作 事	2000/4/1 12:00 ~ 2000/4/1 12:15	20%	12%	4%	...	50%	...
	2000/4/1 12:15 ~ 2000/4/1 12:30	15%	10%	7%	...	40%	...
	2000/4/1 12:30 ~ 2000/4/1 12:45	16%	9%	5%	...	43%	...
	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	⋮	⋮	⋮	⋮	⋮	⋮	⋮
前回累積移動時間取得時刻: 2001/4/12 18:15		記憶装置移動情報					

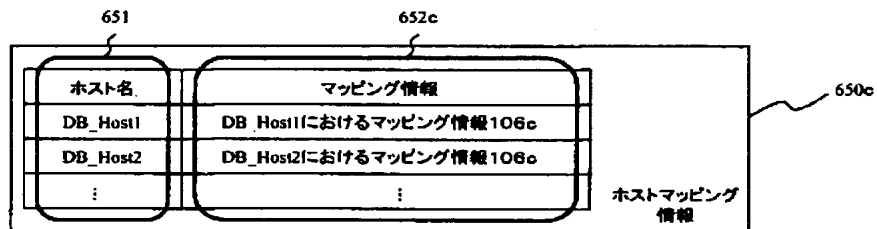
【図 35】

図35

583	602	603	604c
記憶装置名	移動機能情報	データキャッシュ容量	記憶装置ファイル物理記憶位置情報
NAS0	有	4096MB	NAS0におけるファイル物理記憶位置情報510c
NAS1	なし	4096MB	NAS1におけるファイル物理記憶位置情報510c
⋮	⋮	⋮	⋮
記憶装置構成情報			

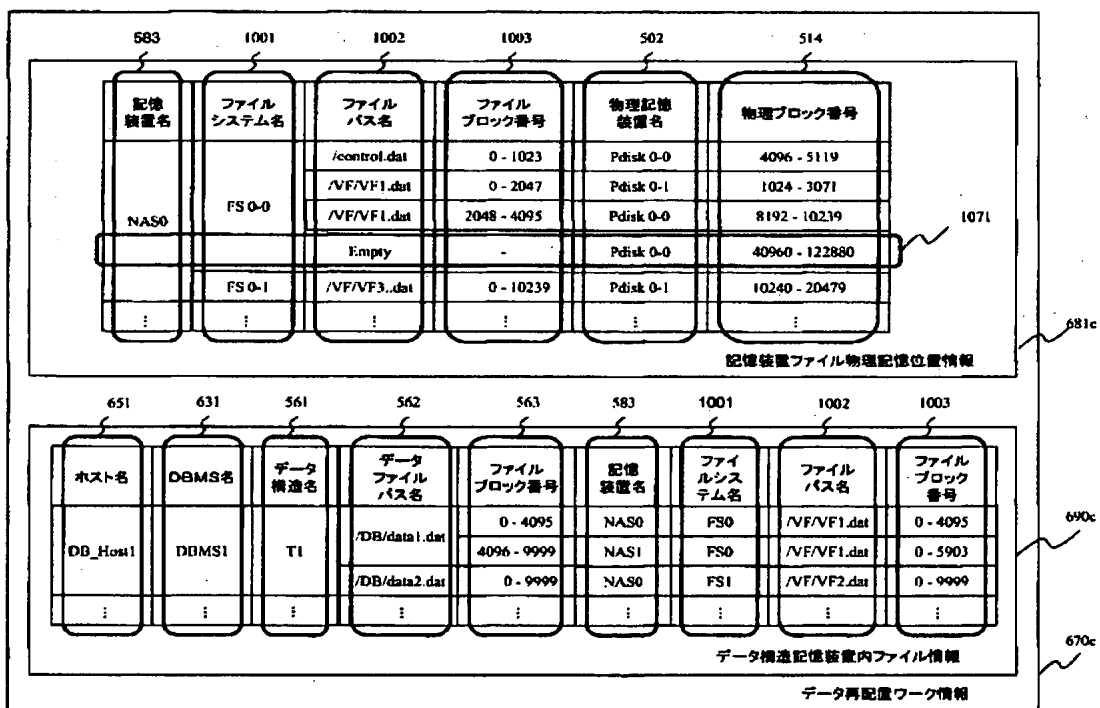
【図 36】

図36



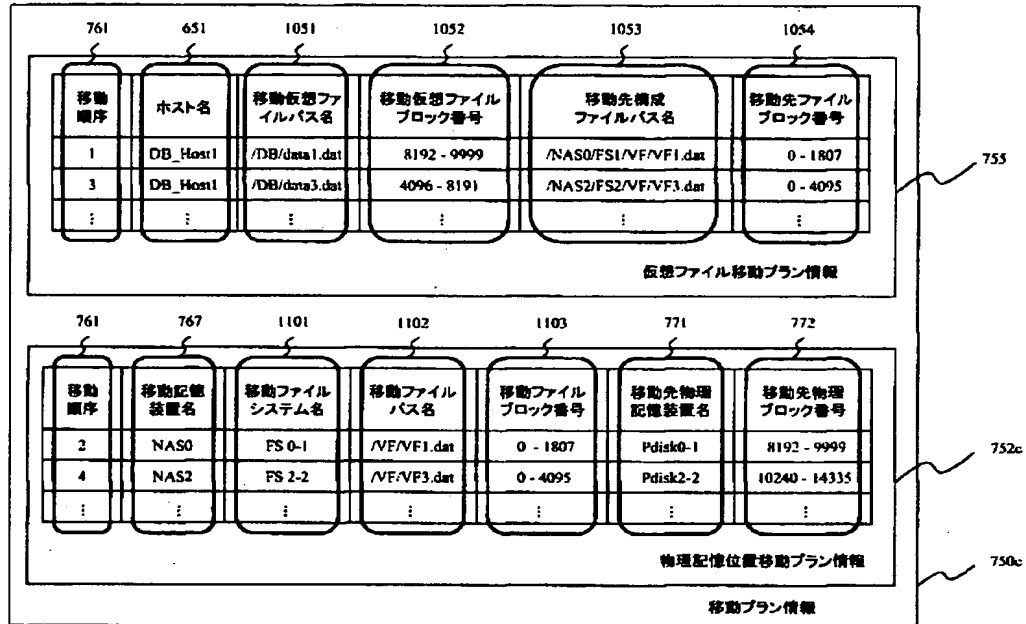
【図 37】

図37



【図 38】

図38



【書類名】 要約書

【要約】

【課題】

データベース管理システム（DBMS）の特性を考慮したデータの再配置を行うことにより、記憶装置からDBMSに良好なアクセス性能を取得する。

【解決手段】

データ記憶位置管理主プログラムは記憶装置からボリューム物理記憶位置管理情報と物理記憶装置稼動情報を収集し記憶装置構成情報と記憶装置稼動情報として記憶し、DBMS内のスキーマ情報から必要な情報を収集してDBMSスキーマ情報として記憶し、ホストにおけるマッピング情報と仮想ボリュームスイッチ中の仮想ボリューム情報を収集しデータ記憶位置管理情報として記憶し、ホストから実行履歴情報を取得してDBMS実行履歴情報として記憶し、それらの情報をもとにより良好な性能特性を持つデータ再配置案を決定し、それを実現するためのデータ移動指示を発行する。

【選択図】 図 1

認定・付加情報

特許出願の番号	特願2001-345523
受付番号	50101661463
書類名	特許願
担当官	第七担当上席 0096
作成日	平成13年11月13日

<認定情報・付加情報>

【提出日】	平成13年11月12日
-------	-------------

出願人履歴情報

識別番号 [000005108]

1. 変更年月日 1990年 8月31日

[変更理由] 新規登録

住所 東京都千代田区神田駿河台4丁目6番地

氏名 株式会社日立製作所